9. Diskurslinguistik und Korpora

Bubenhofer, Noah

DOI: https://doi.org/10.1515/9783110296075-009

Noah Bubenhofer

Diskurslinguistik und Korpora

Der Beitrag skizziert die grundsätzliche korpuslinguistische Perspektive auf Diskurse. Zunächst wird diese Perspektive genauer spezifiziert als Fokussierung auf die Musterhaftigkeit von Sprache und Repräsentation von Daten in Vektorräumen. Danach werden die grundlegenden Schritte des Korpusaufbaus und die Möglichkeiten der Annotation der Daten diskutiert. Schließlich werden die wichtigsten Analysekategorien, die für diskurslinguistische Fragestellungen relevant sind, beschrieben: Kollokationen, Mehrworteinheiten, Keywords, Topic Models und Netzwerkanalysen. Der Beitrag schließt mit Überlegungen, die Korpuslinguistik nicht als Hilfswissenschaft für die Diskursanalyse zu sehen, sondern als Schlüssel zu einem neuen Verständnis des Umgangs mit Daten in den Geisteswissenschaften. In diesem Zusammenhang werden fünf Desiderate genannt, die im Rahmen einer sozial und kulturwissenschaftlich interessierten maschinellen Textanalyse verfolgt werden sollten.

Korpuslinguistik, Korpus, digitale Daten, Annotation, Sprachgebrauchsmuster, XML, Kollokationen, Topic Models, Netzwerke, maschinelle Textanalyse

1 Einleitung

Diskurslinguistische Arbeiten verwenden als empirische Untersuchungsbasis Korpora gesprochener oder geschriebener Sprache, die anschließend auf unterschiedlichen Wegen ausgewertet werden. Es existiert eine längere Diskussion darüber, ob und warum korpuslinguistische Ansätze gewinnbringend für die Diskurslinguistik sind (Busse/Teubert 1994; Spitzmüller/Warnke 2011; Bubenhofer 2009; Teubert 2006) – diese Diskussion soll an dieser Stelle nicht geführt werden. Plädoyers dafür, dass die korpuslinguistische Perspektive für die Diskurslinguistik mehr als ein Werkzeugkasten ist, finden sich in einer Reihe von Publikationen (Bubenhofer 2013; Bubenhofer et al. 2014; Mautner 2012; Scharloth et al. 2013;

Storjohann/Schröter 2011). Stattdessen zielt dieser Beitrag darauf, grundlegende Methoden der korpuslinguistischen Diskursanalyse wiederzugeben und Hinweise für die Aufbereitung und Analyse von Korpora für diskurslinguistische Fragestellungen anzubieten. Damit verbunden ist aber auch eine kritische Reflexion über die sozial- und kulturwissenschaftliche Arbeit mit digitalen Daten und digitalen Methoden.

Hier wird Korpuslinguistik als Zugang verstanden, der Methoden der Computerlinguistik, des Text Minings, der Statistik, visueller Analysemethoden und ähnlicher Felder nutzt, um sozial- und kulturwissenschaftliche Fragestellungen zu untersuchen. An anderer Stelle haben wir diesen Zugang als korpus*pragmatisch* definiert (Scharloth/Bubenhofer 2011; Bubenhofer/Scharloth 2013b, 2015) und argumentiert, dass damit die Korpuslinguistik mehr ist als ein Methodenapparat oder Werkzeugkasten, sondern ein Zugang, der eine neue Sichtweise auf Sprachgebrauch einnimmt: 1) Die sprachliche Oberfläche (Feilke 1996; Feilke/Linke 2009) ist der Ausgangspunkt von Analysen, die 2) im Wechsel datengeleitet, hypothesengenerierend („corpus driven“: Tognini-Bonelli 2001) und hypothesengeleitet 3) auf der Basis von großen Textmengen vorgenommen werden.

Nach wie vor besteht aber ein Desiderat, die theoretischen und methodischen Implikationen korpuslinguistischer Ansätze vor dem Hintergrund des digitalen Zeitalters kritisch zu reflektieren (Bubenhofer/Scharloth 2015). Mit der tiefgreifenden und beinahe flächendeckenden Verdatung der Welt (einerseits über die fortschreitende Digitalisierung analoger Quellen, andererseits über die Masse der ‚digital born‘ Quellen), werden sprachliche Artefakte zählbar. Wenn alle Daten digital vorliegen und unterschiedlichste Informationstypen numerisch repräsentiert sind, können sie deswegen auch beliebig kombiniert werden – ‚Big Data‘ ist damit die mögliche Grundlage für ganz unterschiedliche Formen von Analysen. Als Drittes folgt daraus eine zunehmende Emanzipation der Daten vom Zweck ihrer Produktion, da Daten nicht mehr in Archiven lagern, die nur bestimmte Suchen erlaubten (etwa eine Bibliothek, in der Bücher gefunden werden können), sondern die Daten in ihrer strukturierten, digitalen Form auf mannigfache Weise abfragbar sind (z.B. Kollokationsprofile zu einem Suchausdruck in Abhängigkeit vom Publikationsdatum in allen Büchern mehrerer Bibliotheken). Mit welchen Methoden die Daten dereinst abgefragt werden, kann und muss bei ihrer Aufbereitung nicht erschöpfend definiert werden.

Im Folgenden werden die Schritte der Korpuserstellung, Korpusaufbereitung, Annotation und möglicher grundlegender Analysemethoden beschrieben. Die einsetzbaren Tools und Softwarepakete verändern sich naturgemäß ständig, deshalb werden die Schritte möglichst generisch beschrieben – das Online-Handbuch „Einführung in die Korpuslinguistik“ (www.bubenhofer.com/korpuslinguistik/ – Bubenhofer 2006-2016) bietet diesbezüglich aktuelle Informationen.

Der Fokus dieses Beitrags liegt auf geschriebener Sprache. Ein Teil der Arbeitsschritte und Analysemethoden können grundsätzlich auch für Korpora

gesprochener Sprache eingesetzt werden, trotzdem ergeben sich dort eine Reihe weiterer Probleme, aber auch Chancen, auf die hier nicht eingegangen werden kann.

2 Analyseperspektiven

Wenn Diskurse anhand größerer Datenmengen analysiert werden sollen, stellt sich die Frage nach der Analyseperspektive, die eingenommen werden soll. Textkorpora können z. B. als Textdatenbank oder als Zettelkasten aufgefasst werden. Eine Textdatenbank erlaubt primär eine effiziente Verwaltung von mit Metadaten versehenen Texten, um leicht Texte eines bestimmten Datums, einer Quelle o.ä. zu finden. Damit gewinnt man aber noch keine linguistische Perspektive auf den Text.

Das ist beim Zettelkasten anders, der die Texteinheit durchbricht und den Fokus auf Verwendungsweisen einer beliebigen sprachlichen Einheit, beispielsweise eines Lemmas, lenkt, indem die Verwendungsweisen dieser sprachlichen Einheit übersichtlich aufgeführt und damit systematisch analysierbar gemacht werden.

Der Erfolg der Korpuslinguistik liegt zunächst darin begründet, genau dafür Suchmöglichkeiten und Darstellungsweisen entwickelt zu haben, nämlich die sog. Keyword in Context-Darstellung (KWiC) als Ergebnis einer Suche in großen digitalen Textsammlungen. Dies ist eine triviale Einsicht, jedoch eine relevante, denn lange lag das Interesse der Korpuslinguistik (und deren Nutzerinnen und Nutzer) darin, solche KWiC-Darstellungen zusammenzufassen: Kollokationen, syntagmatische Muster, Distribution von Treffermengen. Dies sind wichtige und erfolgreich angewandte Methoden, die in neuerer Zeit jedoch ergänzt – und vielleicht abgelöst werden durch eine gänzlich andere Perspektive, nämlich jene, die Texte, Wörter oder andere sprachliche Einheiten und deren Eigenschaften als Daten in einem sog. Vektorraum auffassen. Diese beiden unterschiedlichen Perspektiven möchte ich im Folgenden ausführen.

2.1 Syntagmatische Muster und Verteilungen

Elementare Einsicht für die meisten diskurslinguistischen Arbeiten, die Korpora als Grundlage verwenden, ist folgende: In *Mustern* des Sprachgebrauchs spiegeln sich Diskurse wider. Sprachgebrauchsmuster (Bubenhofer 2009) sind demnach rekurrente sprachliche Strukturen, wie etwa Kovorkommen von 1) sprachlichen Ausdrücken (z.B. Lexemen) in gleichen Kontexten oder von 2) sprachlichen Ausdrücken in Texten bestimmter Autorinnen und Autoren, Institutionen, Zeiträumen, Textsorten und dergleichen.

Es gibt verschiedene Möglichkeiten, Kovorkommen von sprachlichen Ausdrücken (1) zu operationalisieren. Letztlich geht es darum, KWiC-Listen von sehr vielen Belegen eines Suchausdrucks zusammenzufassen zu einer Liste von den typischsten *unterschiedlichen* Verwendungsweisen. Die häufigste Operationalisierung ist die Berechnung von Kollokationen (Evert 2009), bei der grundsätzlich in einem Korpus statistisch gesehen auffällig oft zusammen vorkommende Lexeme berechnet werden. Dahinter stecken unterschiedliche statistische Testverfahren, die die Überzufälligkeit der Assoziation bewerten (vgl. grundlegend Evert 2009; Lemnitzer/Zinsmeister 2006; im Kontext diskurslinguistischer Fragestellungen: Bubenhofer 2015). Kollokationsprofile von Lexemen werden für sehr unterschiedliche linguistische Zwecke verwendet, etwa in der Lexikographie und Semantik (Bedeutungskomponenten datengeleitet erarbeiten) oder eben auch in der Diskurslinguistik. Hier ist es oftmals von Interesse, von bestimmten Lexemen die Kollokationsprofile auf der Basis unterschiedlicher Korpora, oder aber die Kollokationsprofile auf der gleichen Datenbasis, aber zwischen unterschiedlichen Lexemen zu vergleichen. Weiter unten werden diesbezüglich Anwendungsbeispiele genannt.

Die Korrelation von sprachlichen Ausdrücken mit beliebigen weiteren Texteigenschaften wird über unterschiedliche distributionelle Analysen ergründet (2). Im einfachsten Fall werden beispielsweise Frequenzen eines Suchbegriffs in unterschiedlichen Teilkorpora, z.B. in einem nach Publikationsdatum der Texte diachron nach Jahren o.ä. gegliederten Korpus, untersucht. Oder die Verwendungsfrequenzen eines grammatischen Phänomens (z.B. Passivformen) in Zeitungskorpora zu unterschiedlichen Themen.

Dabei existiert aus korpuslinguistischer Perspektive eine Diskussion darüber, welches geeignete Maße sind, um die Frequenzen zu messen und zu vergleichen; dass das Maß in Relation zur jeweiligen Korpusgröße stehen muss, versteht sich von selbst. Darüber hinaus wurden aber auch Maße wie ‚Frequenzklassen' vorgeschlagen, die robuster gegenüber unterschiedlichen Korpusgrößen sind und deshalb relativen Maßen wie ‚Treffer pro Mio. laufender Wortformen' o.ä. vorgezogen werden (vgl. für eine ausführliche Diskussion Perkuhn et al. 2012, 78). Weiter können wiederum statistische Testverfahren eingesetzt werden, um zu messen, ob die gemessenen Frequenzunterschiede signifikant sind oder nicht (Kilgarriff 2001).

All diesen Methoden gemeinsam ist die Tatsache, dass die Abhängigkeit einer Variable von einer gegebenen, unabhängigen Variable gemessen wird. Wenn ich allerdings so herausfinde, dass ein Lexem X in einem Korpus A signifikant häufiger vorkommt als in B, oder wenn ich herausfinde, dass ein Lexem Y signifikant häufig zusammen mit X vorkommt, teste ich nur genau diese Zusammenhänge und keine weiteren (außer ich wiederhole die Messung systematisch mehrmals mit anderen unabhängigen Variablen). In Ansätzen, die unter Labeln wie ‚Data Mining' in ‚Big Data' o.ä. geführt werden, versucht man jedoch normalerweise

komplexere Abhängigkeiten zu testen. Dafür müssen Texte als Daten mit Eigenschaften in Vektorräumen modelliert werden.

2.2 Daten im Vektorraum

Wenn vor einem diskursanalytischen Hintergrund die Frage interessiert, welche Themenschwerpunkte in einem Diskurs vorherrschend sind und durch welches Vokabular diese geprägt werden, explodiert die Anzahl der zu berücksichtigenden Variablen: Auf Lexemebene gedacht ist jeder in den Texten vorkommende Lexemtype mit den Frequenzen in den jeweiligen Texten eine Variable. Es wäre nun aufwändig, für jeden Lexemtype die Frequenzen in jedem Text zusammenzutragen – und auch dann wäre unklar, welche nun tatsächlich für bestimmte Texte bedeutend wären oder ob für bestimmte Gruppen von Texten eine für die jeweilige Gruppe spezifische Kombination von Lexemen existiert, die für einen thematischen Cluster stehen.

Um solche Fragen zu verfolgen, werden Daten als Sammlung von Objekten mit Eigenschaften in einem Vektorraum angesehen: Ein Datensatz (ein Korpus) besteht aus Objekten (Texten). Diese Objekte weisen bestimmte Eigenschaften (‚Features') auf: Darin vorkommende Lexeme, andere sprachliche Phänomene, bestimmte Metadaten etc. Diese Eigenschaften werden nun in einer systematischen Art und Weise gemessen und als Matrix repräsentiert: Jedes Objekt nimmt eine Zeile ein, jede Eigenschaft eine Spalte, sodass für jedes Objekt die Ausprägung dieser Eigenschaft notiert werden kann. Das Ergebnis ist eine Tabelle mit Werten des gesamten Datensatzes, wie wir sie auch manuell erstellen würden.

Text	fX	fY
A	15	28
B	25	11
C	27	9
D	30	10
E	30	30
F	31	10
G	28	28
H	27	30

Diese Tabellen können sehr breit sein, also viele Spalten (Eigenschaften) ausweisen. Jede Eigenschaft, die irgendwo in einem der Objekte auftritt und die uns interessiert, die wir also messen, muss aufgeführt werden, wobei wir für jedes Objekt einen Messwert haben (kommt vor, kommt gar nicht vor, ggf. wie häufig oder mit welcher Assoziation). Die Folge der Ausprägungen aller Eigenschaften wird nun Vektor genannt; es handelt sich um eine Gerade in einem mehrdimensionalen Raum. Das Potenzial, was sich daraus ergibt, lässt sich besser erklären, wenn man zunächst von einem einfacheren Datensatz ausgeht: Nehmen wir an, für unsere Objekte erfassen wir nur die Frequenzen zweier Lexeme X und Y. Wir können nun die Werte für X und Y in einem Diagramm mit den zwei Achsen X und Y eintragen und erhalten ein Streudiagramm (vgl. Abbildung 1).

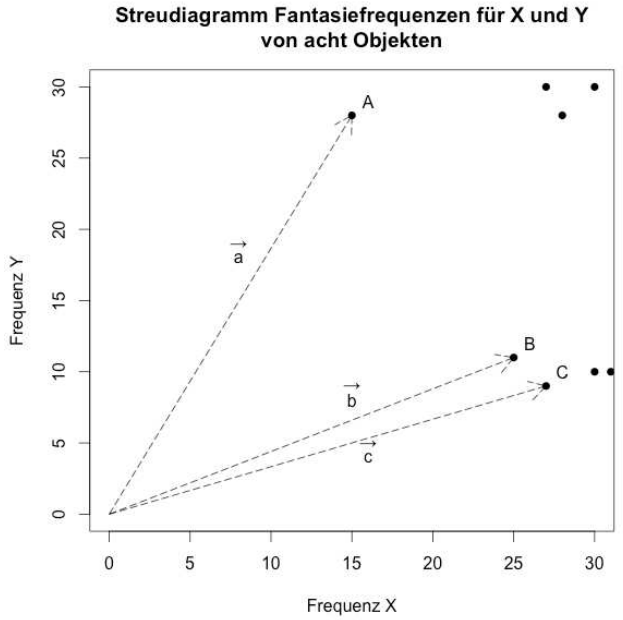


Abbildung 1

Je nachdem, wo die Punkte liegen, können wir die Objekte charakterisieren: Liegen die Punkte im Bereich oben rechts, handelt es sich um Objekte, in denen beide Lexeme häufig vorkommen. Punkte in der rechten Ecke unten charakterisieren Objekte, in denen das Lexem X vorherrschend ist, Y aber kaum vorkommt usw.

Wir fassen nun die X- und Y-Werte als Vektoren auf: Der Vektor, die Gerade, beginnt beim Schnittpunkt 0,0 der beiden Achsen X und Y und endet beim jeweiligen Punkt des Objekts. Dieser Vektor beschreibt die Eigenschaft des

Objektes geometrisch im Vektorraum. Wenn die Vektoren zweier Objekte in die ähnliche Richtung zeigen (in Abbildung 1 Vektoren der Objekte B und C), ähneln sich die beiden Objekte, ansonsten unterscheiden sie sich mehr oder weniger (A unterscheidet sich stark von B und C). Das Maß der Abweichung kann über geometrische Maße gemessen werden, etwa indem die Distanz zwischen den Punkten (euklidische Distanz) oder der Winkel gemessen wird, den die beiden Vektoren aufmachen (Kosinus-Distanz).

Wir können uns zu den beiden Dimensionen X und Y noch eine dritte Dimension Z vorstellen, wenn wir die Frequenz eines weiteren Lexems erfassen. Dies lässt sich noch immer visualisieren, indem ein dreidimensionales Streudiagramm gezeichnet wird und zusätzlich die Position auf der in den Raum zeigenden Z-Achse verortet wird. Wenn wir weitere Eigenschaften hinzufügen, bewegen wir uns entsprechend in einem n-dimensionalen Raum, der sich nicht mehr visualisieren, aber nach wie vor rechnerisch behandeln lässt. Jedes Objekt wird nach wie vor durch einen Vektor in einem mehrdimensionalen Raum repräsentiert und die Ähnlichkeit oder Differenzen zwischen den Objekten lässt sich noch immer über geometrische Operationen berechnen. Darin liegt das Potenzial begründet, wenn die Komplexität von Daten in einem mehrdimensionalen Raum repräsentiert und damit operiert werden kann. Das geisteswissenschaftliche Vergleichen von Dingen wird damit zu einem geometrischen Messen von Distanzen im Vektorraum: „The concept of a geometric feature space allows us to take the most basic method of humanities – a comparision – and extend it to big cultural data. In the same time, it allows us (or forces us, if you prefer) to quantify the concept of difference“ (Manovich 2015, 26).

Aus diskurslinguistischer – und generell geisteswissenschaftlicher Sicht bedeutet dies: Große Datenmengen sind gut handhabbar und reiche, differenzierte Annotationen gut nutzbar. Zusätzlich kommt hinzu, dass diese Methoden neue Analysestile ermöglichen: Explorative, datengeleitete Zugänge sind möglich, ggf. im Verbund mit hypothesengeleiteten und/oder auch qualitativen Zugängen (Bubenhofer et al. 2014; Bubenhofer/Scharloth 2015).

Sobald die Eigenschaften von Objekten als Vektoren repräsentiert werden, können eine Vielzahl von Algorithmen darauf angewandt werden. So versuchen Clustering-Methoden Gruppen von Objekten zu identifizieren, die ähnliche Eigenschaften aufweisen und sich von anderen Objekten (und Gruppen) möglichst stark unterscheiden. Bei einem zwei- (wie in Abbildung 1) oder dreidimensionalen Streudiagramm *sehen* wir diese Gruppen in Form von nahe beieinander stehenden Punkten. Bei mehr Dimensionen jedoch nicht. Die vom Algorithmus identifizierten Cluster können anschließend von der Forscherin als Kategorien gedeutet werden und führen allenfalls sogar zu Kategorisierungen der Daten, an die man vorher gar nicht gedacht hat. Solche explorativen, sog. ‚unüberwachten‘ Methoden kontrastieren mit ‚überwachten‘ Lernmethoden, wo bereits manuell oder anderweitig klassifizierte Daten die Grundlage sind, um daraus algorithmisch ein

statistisches Modell zu erstellen, mit dem neue, nicht-klassifizierte Daten maschinell klassifiziert werden können (Carstensen et al. 2010, 591; Manovich 2015, 24).

Manovich argumentiert im Kontext der Digital Humanities, dass solche Klassifikationsaufgaben, also überwachte Lernverfahren, die darauf hinzielen, bestehende Kategoriensysteme für eine maschinelle Klassifikation von großen Datenmengen zu nutzen, nicht sehr interessant ist: „Why should we use computers to classify cultural artifacts, phenomena or activities into a small number of categories? Why not instead use computational methods to question the categories we already have, generate new ones, or create new cultural maps that relate cultural artifacts in original ways?“ (Manovich 2015, 24) Unüberwachte, datengeleitete Methoden erlauben stattdessen einen explorativen Zugang zu riesigen Datenmengen, was gerade auch für diskurslinguistische Fragestellungen interessant ist (Scharloth et al. 2013).

3 Korpuserstellung

3.1 Textauswahl

Ein zu untersuchender Diskurs ist nicht zwingend durch ein klar definiertes Korpus repräsentierbar, da Diskurse als Aussagensysteme begriffen werden müssen, die quer zu Texten liegen können (Spitzmüller/Warnke 2011, 25, 88, 91). Trotzdem ist es aus forschungspraktischen Gründen unumgänglich, eine Auswahl an Texten für ein Korpus zu treffen. Folgende Strategien können aber dem berechtigten Einwand zu enger Korpusgrenzen entgegenkommen (vgl. ausführlich Bubenhofer 2009):

> 1) Dank immer besserer Verfügbarkeit digitaler Texte und größerer Speicher- und Rechenkapazitäten ist es oft möglich, mit sehr großen Textkorpora zu arbeiten, die zunächst gar keiner thematischen Eingrenzung unterworfen sind. Um beispielsweise massenmediale Krisendiskurse in Nachkriegsdeutschland zu untersuchen, stehen mit den digitalen Archiven der Wochenzeitschrift ‚die Zeit‘ und des Magazins ‚der Spiegel‘ Quellen zur Verfügung, mit denen ein Korpus aller in Nachkriegsdeutschland bis heute publizierten Artikel aufgebaut werden kann. Bei der späteren Analyse könnten immer wieder neue Eingrenzungen (thematisch, zeitlich, bezüglich Textsorten etc.) vorgenommen werden. Auch große öffentlich verfügbare Korpora wie das DeReKo (Kupietz et al. 2010) enthalten genug große Textmengen verschiedener Typen, um Korpora zu bilden, die zunächst thematisch überhaupt nicht eingegrenzt werden. Schwierig bis unmöglich ist dieses Vorgehen selbstverständlich bei schwieriger Quellenlage.

2) Der Vorteil korpuslinguistischer Zugänge zu Texten liegt darin, dass die Texteinheit oft gar keine Rolle spielt: Die typische Key-Word in Context-Ansicht (KWiC) bricht die Textgrenzen auf und lenkt die Aufmerksamkeit auf Aussagemuster, die über Textgrenzen hinaus gehen. Trotzdem sind bei Bedarf die Einzeltext-Informationen (Metadaten, Artikelgrenzen) verfügbar.

Bei der Auswahl eines Korpus für eine diskurslinguistische Analyse sollte idealerweise zunächst nur definiert werden, welche Sprachdomänen, medialen Erscheinungsformen, Textsorten oder zeitlichen Perioden Grundlage für die Analysen sind. Diese Definition orientiert sich verständlicherweise auch an den praktischen Möglichkeiten, die durch technische, finanzielle und nicht zuletzt juristische (Perkuhn et al. 2012, 52) Grenzen bestimmt werden. Aus dieser Grundgesamtheit wird dann eine Stichprobe gezogen, die entweder zufällig oder aber geschichtet ist, also bezüglich bestimmter Dimensionen, die für die Analyse als entscheidend angesehen werden, ausgeglichen ist. Eine solche Zufallsstichprobe hat das Ziel, eine für die Grundgesamtheit repräsentative Textmenge zusammenzustellen, sodass Beobachtungen in der Stichprobe auf die Grundgesamtheit interpoliert werden können.

Während dieses Vorgehen bei sozialwissenschaftlichen empirischen Untersuchungen hochgradig standardisiert ist, stellen sich in der Korpuslinguistik eine Reihe von Problemen (Perkuhn et al. 2012, 46). Bei Korpora, die ‚die deutsche Sprache' repräsentieren sollen, ist offensichtlich, dass dieses Ziel nicht erreicht werden kann, ohne vorher genau definiert zu haben, was zur ‚deutschen Sprache' gehört – was ohne ganz deutliche Eingrenzungen unmöglich ist. Aber wie verhält es sich mit Diskursen? Migrationsdiskurse zeigen sich nicht nur in der massenmedialen Politik-Berichterstattung, sondern auch in anderen Themen (z.B. Ökologie: Einwanderung von ‚fremdem' Saatgut) und in anderen medialen Formen (Jung 1996). Um eine Grundgesamtheit zu definieren, sind Eingrenzungen notwendig – aber nicht zwingend thematischer Art. Beispielsweise könnte für das Beispiel Migrationsdiskurs die Grundgesamtheit auf deutschsprachige Pressetexte und Webforen-Diskussionen in einem bestimmten Zeitraum beschränkt werden. Mit der (geschichteten) Stichprobe würde man sich auf bestimmte Pressetitel und Foren beschränken, dort aber jeweils alle Texte des definierten Zeitraums ins Korpus integrieren (Vollerhebung).

3.2 Untersuchungs- und Referenzkorpus

Einige der im Folgenden vorgestellten Analysemethoden beruhen auf Korpusvergleichen: Mit Keyword-Analysen wird beispielsweise das typische Vokabular eines Korpus im Vergleich zu einem Referenzkorpus berechnet. Das Referenzkorpus dient also dazu, die Besonderheiten des Untersuchungskorpus hervorzuheben.

Meist ist es nicht sinnvoll, gleich zu Beginn eines Projekts ein festes Referenzkorpus zu definieren. Denn die Wahl des Referenzkorpus entscheidet darüber, welche Parameter (Thema, Textsorte, Zeitperiode etc.) miteinander verglichen werden können; idealerweise unterscheiden sich Untersuchungs- und Referenzkorpus jeweils bezüglich eines Parameters, um die beobachteten Unterschiede als Effekt dieses Parameters erklären zu können.

In einer Untersuchung zu Veränderungen von Sprachgebrauchsmustern in der Neuen Zürcher Zeitung von 1995 bis 2005 wurde das Korpus, eine Zufallsstichprobe aller im Zeitraum erschienenen Artikel, nach unterschiedlichen Kriterien unterteilt (Bubenhofer 2009, 197). Beispielsweise wurden zwei Zeiträume 1995-1997 und 2003-2005 definiert und pro Ressort (‚Ausland', ‚Inland', ‚Wirtschaft' etc.) Teilkorpora erstellt. Bei der Gegenüberstellung des Auslands-Korpus 1995-1997 und des Auslands-Korpus 2003-2005 sind alle Parameter (Medium, Textsorten, Ressort, grundsätzlich vorkommende Themen) gleich bis auf den Publikationszeitpunkt. Die bei der Analyse sich ergebenden Unterschiede können demnach als zeitgebundene Veränderungen des Sprachgebrauchs (was indirekt natürlich auch mit unterschiedlichen Themen zusammenhängen kann) interpretiert werden. Zusätzlich könnte aber auch das gesamte NZZ-Korpus als Referenz zu einem Untersuchungskorpus ‚Feuilleton-Artikel' oder ‚Artikel zum Thema X' definiert werden, um die Spezifika von Feuilleton-Artikeln bzw. Artikeln zu einem Thema X herauszuarbeiten.

Als Referenzkorpus kann je nach Anlage auch ein externes Korpus verwendet werden: Das DeReKo (Kupietz et al. 2010) eignet sich mit seiner Größe von 24 Milliarden laufenden Wortformen (Stand 2014) als Referenz, um beispielsweise in der Gegenwartssprache typische Verwendungsweisen bestimmter sprachlicher Ausdrücke zu untersuchen und mit denen des eigenen Untersuchungskorpus zu vergleichen (vgl. dazu ausführlicher Bubenhofer 2013).

Beim Design des Korpus empfiehlt es sich demnach, für das Forschungsvorhaben potenziell interessante Korpusvergleiche vorzusehen. Demnach ist eine Vollerhebung bestimmter Quellen (z.B. alle Zeitungsartikel einer Zeitung in einem bestimmten Zeitraum) auch aus Gründen der Referenzkorpus-Bildung von Vorteil.

4 Korpusaufbereitung und Annotation

4.1 Überführung in strukturierte Daten

4.1.1 XML als nachhaltiges Datenformat

Für diskurslinguistische Untersuchungen ist eine reiche Auszeichnung der Daten mit Metadaten sehr bedeutend. Zwar soll es das Korpus ermöglichen, Aussagemuster über Textgrenzen hinweg zu finden, gleichzeitig müssen die Metadaten aber zur Verfügung stehen, um beliebig Teilkorpora zu bilden oder das Streuungsverhalten bestimmter Phänomene über Textsorten, Autorinnen und Autoren, Zeitabschnitte etc. zu untersuchen.

Deswegen ist es unumgänglich, das Korpus in einem strukturierten oder semistrukturierten Format abzulegen. Dafür eignet sich die Auszeichnungssprache XML (Bray et al. o. J.), mit der grundsätzlich ein eigenes, maßgeschneidertes Kategorien- und Auszeichnungssystem umgesetzt werden kann. Im Folgenden sei ein einfaches Beschreibungsmodell der Metadaten gegeben, das beliebig erweitert werden kann:

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
	<document id="imo345">
		<header>
			<title>Titel des Textes</title>
			<author>Autor</author>
			<date>2012-04-01</date>
		</header>
		<body>
			Hier folgt der Text...
		</body>
	</document>
</corpus>
```

Es existieren allerdings XML-Standards zur Codierung von Textkorpora. Dazu gehören die Standards TEI (‚Text Encoding Initiative' TEI Consortium 2014; Stührenberg 2012) oder XCES (‚Corpus Encoding Standard for XML' Ide et al. 2000), die versuchen, für möglichst viele unterschiedliche Texte Codiervorgaben zu geben. Dazu gehören nicht nur Vorgaben für die Codierung der Metadaten, sondern auch (hauptsächlich bei TEI) die Möglichkeit, Layoutinformationen im Text zu codieren. TEI wird laufend weiterentwickelt und es existieren Hilfsmittel, um aus dem vergleichsweise komplexen Standard vereinfachte Versionen für eigene Projekte abzuleiten.

Die XML-Auszeichnungen folgen einem Regelapparat, der die zu verwendenden Elemente und die hierarchische Struktur, nach denen sie angeordnet werden können, beschreibt (Means 2004). Es gibt zwei Möglichkeiten, diesen Regelapparat zu beschreiben: Die ältere Variante ist das Verfassen einer DTD (Document Type Definition) – neuer, und der DTD überlegen, da die Regeln genauer beschrieben werden können, ist das XML-Schema (van der Vlist 2011).

Folgt man dem TEI-Standard, kann man auch das entsprechende XML-Schema verwenden und XML-Dokumente, die man selber erstellt, gegenüber dem Schema validieren lassen. Dazu dienen beispielsweise XML-Editoren oder ein XML-Parser.

Ein gewichtiger Vorteil von XML liegt darin, dass Dokumente, die sich bereits in einer XML-Struktur befinden, vergleichsweise einfach in eine andere XML-Struktur konvertiert werden können. So kann das weiter oben dargestellte einfache XML-Schema in XML-TEI konvertiert werden, wobei natürlich ggf. Informationen ergänzt werden müssen. Damit ist XML ein Format, das sich für eine nachhaltige Archivierung von Daten eignet.

Um XML-Dokumente eines bestimmten Schemas in ein XML-Dokument eines anderen Schemas zu überführen, wird ein XSL-Stylesheet geschrieben (XSLT, Tidwell 2008). Dieses regelt, welche XML-Elemente in bestimmte andere XML-Elemente überführt werden sollen. Mit einem solchen Stylesheet können dann beliebig viele XML-Dokumente, die dem gleichen Schema folgen, in XML-Dokumente eines anderen Schemas überführt werden.

4.1.2 Erstellung von XML-Dokumenten

XML-Dokumente können über verschiedene Wege erstellt werden. Die Wahl des Weges hängt davon ab, in welchem Ausgangsformat sich die zu konvertierenden Dokumente befinden. In aller Kürze kann dazu Folgendes zusammengefasst werden:

- Ausgangsdokumente sind in einem **strukturierten Format** vorhanden, z.B. als Datenbanktabelle, CSV-Datei (kommaseparierte Datei, etwa als Export aus einer Tabellenkalkulations-Software wie Microsoft Excel oder OpenOffice Calc) etc.: Eine Überführung in XML ist besonders einfach und oft direkt aus der entsprechenden Software hinaus möglich. Ggf. muss das dabei entstandene XML-Format in das gewünschte Format, z.B. TEI, mittels XSLT konvertiert werden.

- Ausgangsdokumente sind in einem **semi-strukturierten Format** vorhanden, z.B. als XML oder HTML-Dokumente: XML-Dokumente können leicht mittels XSLT in das gewünschte XML-Schema überführt werden. Ein Spezialfall sind HTML-Dokumente: HTML gehört zur Familie der XML-Dokumente, gehorcht aber nicht den strengen Regeln der Wohlgeformtheit wie XML (so müssen z.B. sich öffnende Elemente nicht zwingend geschlossen werden – ein Browser interpretiert solche HTML-Dokumente großzügig). Daher ist vor der Weiterverarbeitung mit XSLT die Überführung in XHTML notwendig. Dafür existieren Konverter, die zudem typische Syntaxfehler in HTML-Dokumenten korrigieren und XML-konformes XHTML ausgeben.

- Ausgangsdokumente sind in einem **unstrukturierten Format** vorhanden, z.B. OCR-erkannter Text (Text, der von einer Optical Character Recognition-Software aus Bilddaten erkannt worden ist), PDFs, Word-Dokumente etc.: Bei solchen Dokumenten ist eine Überführung in XML mit erheblich mehr Aufwand verbunden. Entweder muss dabei komplett manuell gearbeitet werden oder aber es können Heuristiken entwickelt werden, mit denen Strukturen in den Dokumenten (z.B. Titel, Untertitel, Datumsangabe etc.) erkannt und in entsprechende XML-Strukturen überführt werden. Dazu können beispielsweise Scriptsprachen verwendet werden, mit denen solche Heuristiken programmiert werden.

Je nach Ausgangslage muss also mehr oder weniger Aufwand in die Konvertierung des Materials in XML einberechnet werden.

4.2 Linguistische Annotation der Textdaten

Mit dem Annotieren von Textdaten ist das Hinzufügen von beliebigen linguistischen Informationen zu den Wortformen oder Gruppen von Wortformen gemeint (Lemnitzer/Zinsmeister 2006; Perkuhn et al. 2012, 57). Typischerweise werden mittels eines Part-of-Speech-Taggers automatisch morphosyntaktische Informationen (Wortartklassen) hinzugefügt sowie die jeweilige Grundform (Lemmatisierung, Stemming). Weiter können aber auch syntaktische Informationen (Phrasentypen, Satzglieder o.ä.) annotiert werden sowie zahlreiche weitere Kategorien.

Bei großen Textkorpora versucht man diese Annotationen maschinell zu erzeugen. Morphosyntaktische Tagger annotieren mit relativ hoher Sicherheit auf der Basis eines Lexikons und statistischer Modelle, die typische Wortarten-Kombinationen anhand eines manuell annotierten Trainingskorpus gelernt haben.

Neben maschinellen Methoden der Annotation können Textdaten jedoch auch manuell oder halbautomatisch annotiert werden. Oft sprechen forschungspraktische Gründe dagegen, große Textmengen manuell zu annotieren, um komplexe Phänomene zu annotieren, ist ein manuelles (z.B. in einem Teilkorpus) oder halbautomatisches Annotationsverfahren aber oft unumgänglich. Ich beschränke mich im Folgenden auf eine Übersicht über maschinelle Annotationsmethoden.

Linguistische Diskursanalysen sind, zumindest auf den ersten Blick, weniger an grammatikalischen, sondern eher an lexikalischen Phänomenen interessiert. Deshalb könnte man zum Schluss kommen, dass eine linguistische Annotation der Daten, also beispielsweise das Hinzufügen von Wortartklassen zu den Wortformen, nicht von Interesse ist. Allenfalls wäre eine Lemmatisierung der Wortformen, also das Rückführen der Wortformen auf Grundformen, eine Erleichterung für die Suche im Korpus – doch rechtfertigt dies den Aufwand? Zwei Gründe

sprechen jedoch dafür, zumindest eine morphosyntaktische Annotation und Lemmatisierung (Zinsmeister 2015) vorzunehmen:

- Auch aus diskurslinguistischer Sicht ist die Beobachtung von typischen syntaktischen Mustern (oder Veränderungen davon) in den Daten interessant. Zu den zahlreichen Beispielen gehören die Verwendung von Modalverben (mit oder ohne Negation?) in Verbindung mit typischen Subjekten oder Objekten, Verwendung von Tempusformen, die Anteile von Aktiv- und Passivkonstruktionen, von Partikeln in Verbindung mit bestimmten Nomen (*für* X, *gegen* X, *mit* X etc.), Anzeige typischer Adjektive in Verbindung mit bestimmten Nomen und dergleichen mehr. Daneben sind morphosyntaktische Kategorien bei datengeleiteten Analysen eine wichtige Bereicherung, wie weiter unten gezeigt wird. Aber auch bei Analyseverfahren, die primär ohne morphosyntaktische Annotation auskommen, können diese Informationen in einem zweiten Schritt von großer Hilfe sein – ein Beispiel ist die Berechnung von Kollokationen: Bei annotierten Daten können die Kollokatoren anschließend nach Wortartklassen gefiltert werden, was die Interpretation systematisiert.

- Der Nutzen einer morphosyntaktischen Annotation ist zudem zu einem relativ geringen Preis zu haben: POS-Tagger gehören zu den Standardanwendungen der Computerlinguistik und sind leicht einsetzbar (Beispiele für POS-Tagger sind der ‚TreeTagger' – Schmid 1994; und der ‚RFTagger' – Schmid/Laws 2008). Wenn die Daten in einem XML-Format vorliegen, ist die Annotation technisch relativ problemlos. Einige Korpusmanagement-Systeme bieten zudem Standardannotationen, darunter die morphosyntaktische Annotation, an, ohne zusätzliche Software installieren zu müssen.

Neben morphosyntaktischer Annotation sind weitere maschinell anwendbare Annotationskategorien für linguistische Diskursanalysen von Interesse:

- Automatische syntaktische Analysen versuchen die syntaktischen Strukturen der Sätze zu identifizieren, wobei man Chunk-Parsing (Carstensen et al. 2010, 275) und vollständiges Syntaxparsing (Carstensen et al. 2010, 303) unterscheidet. Chunk-Parsing, auch partielles Parsing genannt, beschränkt sich auf das Erkennen von Phrasen mit inhaltstragenden Wörtern als Kopf, ohne die Abhängigkeiten der einzelnen Chunks untereinander zu bestimmen. Diese Technik ist um einiges robuster und erfolgreicher als das Parsing der kompletten Dependenzstruktur, dafür auch weniger informativ. Syntaktisch annotierte Daten können für diskurslinguistische Fragestellungen interessant sein, wenn grammatische Eigenschaften (Textkomplexität, Topikalisierung, Aktiv-/Passivverwendung etc.) diskursive Bedeutung haben könnten.

- Named Entity Recognition (NER) – Eigennamenerkennung: Es

existieren Systeme, die aufgrund von Wortlisten und statistischem Wissen maschinell Eigennamen erkennen und klassifizieren, indem sie z.B. Personen-, Firmen- und Institutionennamen sowie Toponyme unterscheiden. Bei einem diskursanalytischen Interesse für Akteure sind solche Informationen ohne Zweifel interessant. Auch die Verwendung von Toponymen in Diskursen kann erhellend sein oder die Berechnung von ‚Geokollokationen' – typische Kollokatoren zu Toponymen – deckt die diskursive Prägung von Orten auf (Bubenhofer 2014).

- Bei mehrsprachigen Daten, bei denen nicht bereits Sprachinformationen verfügbar sind, ist wahrscheinlich eine maschinelle Sprachidentifikation sinnvoll. Diese arbeiten beispielsweise mit Sprachmodellen, die auf Häufigkeiten von Buchstaben-N-Grammen (also Folgen von Buchstaben) beruhen und funktioniert für die gängigen Sprachen sehr erfolgreich (Dunning 1994).
- Ebenfalls von Bedeutung sind linguistische Ressourcen wie semantische Lexika (z.B. WordNet: Miller 1995; GermaNet: Kunze/Lemnitzer 2002), die den Wortschatz in sog. Synsets, also der Menge der Synonyme eines Konzeptes, modelliert und semantische Relationen festlegt. Diese Ressourcen können verwendet werden, um beispielsweise Wörtern in einem Korpus Hyperonyme zuzuweisen. Es gibt zahlreiche weitere Ansätze, Wissensbasen aufzubauen, indem etwa die Wikipedia als Grundlage für Extraktionen von Ontologien verwendet wird (Gabrilovich/Markovitch 2006; Zesch et al. 2008).

Aus diskurslinguistischer Sicht ergeben sich zwar interessante Optionen, um Korpusdaten maschinell reich zu annotieren, es stellen sich aber auch Probleme.

Zunächst muss man bei computerlinguistischen Annotationsverfahren grundsätzlich damit leben, dass sie nicht fehlerfrei arbeiten. Auch wenn beispielsweise bei einem POS-Tagger bei Zeitungstexten 95% Akkuratheit erreicht wird – also 95% der Wörter richtig annotiert werden – bedeutet dies bei einer durchschnittlichen Satzlänge von 17 Wörtern, dass ungefähr jeder zweite Satz einen Fehler aufweist (Zinsmeister 2015, 96). Lässt man längere Sätze außen vor, verbessert sich die Rate etwas. Für diskurslinguistische Fragestellungen muss das nicht zwingend problematisch sein, da die Akkuratheit auch nicht gleichmäßig auf die Wortarten verteilt ist. Dies gilt für alle Verfahren, die mit statistischen Modellen arbeiten, die auf der Basis von Trainingskorpora berechnet worden sind. Phänomene, die häufig sind, werden deshalb auch sicherer erkannt, seltenere eher nicht. Möchte man also beispielsweise die Adjektivvariation in einem Korpus untersuchen, werden Annotationsfehler nicht stark ins Gewicht fallen. Bei selteneren Phänomenen oder Textsorten, die sich stark von den Trainingsdaten des Taggers unterscheiden, hingegen schon.

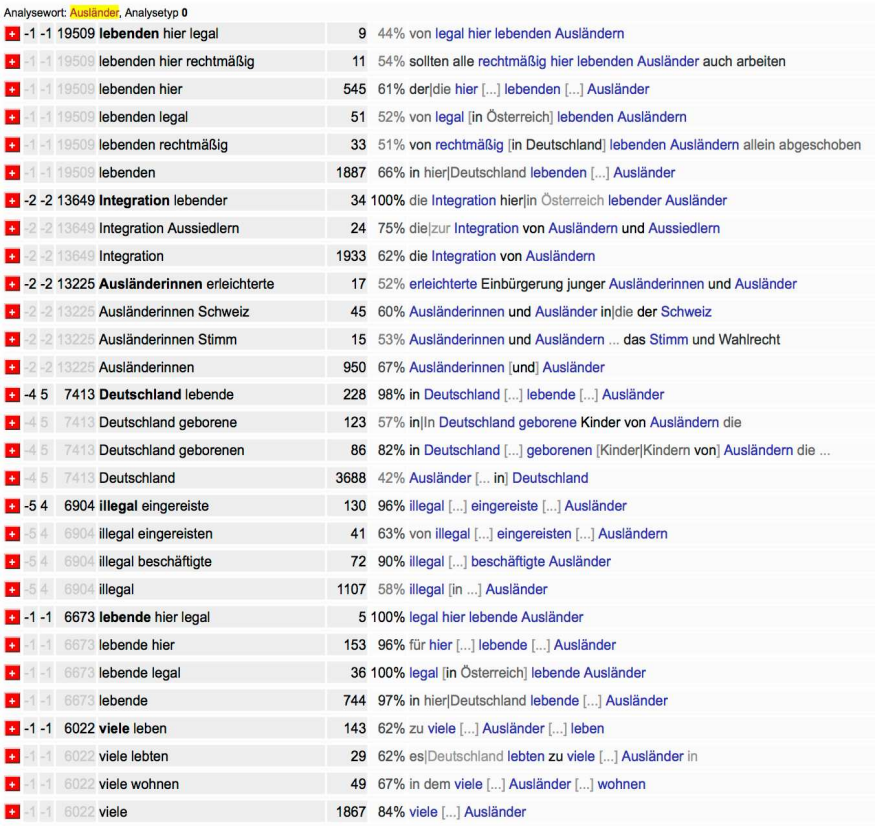
Weiter ist man bei der Verwendung von computerlinguistischen Tools und Ressourcen von Designprinzipien und Prämissen abhängig, die bei der Erstellung

der Tools festgelegt worden sind und die nicht immer transparent und selten veränderbar sind. Ein semantisches Lexikon wie GermaNet geht beispielsweise nicht von einem sich dynamisch verändernden Wortschatz aus, der darüber hinaus in unterschiedlichen Diskursen unterschiedlich verwendet wird. Die automatische Erkennung von Eigennamen ist abhängig von der verwendeten Definition des Konzeptes ‚Eigenname'. Solche Prämissen reproduzieren also Episteme, die über die Diskursanalyse freigelegt werden sollen. Es ist deswegen unabdingbar, Annotationstechnologien informiert und nicht als Black Box anzuwenden und die damit verbundenen Implikationen abschätzen zu können.

5 Analysekategorien und Beispiele

Abschließend möchte ich die Praxis der korpuslinguistischen Diskursanalyse an Beispielen zeigen. Damit konkretisieren sich auch zentrale Analysekategorien, die oben bereits erwähnt und an anderer Stelle (Bubenhofer/Scharloth 2013a, 2015) ausführlich beschrieben werden. Zu diesen zentralen Kategorien gehören (neben anderen):

- Kollokationen/Kookkurrenzen: Statistisch signifikantes Kovorkommen von sprachlichen Ausdrücken, meist Lexemen.
- N-Gramme/Mehrworteinheiten: Folgen von Wortformen, Lemmata, Wortartklassen o.ä. oder Kombinationen davon.
- Keywords: Sprachliche Ausdrücke, die in einem Untersuchungskorpus im Vergleich zu einem Referenzkorpus statistisch signifikant häufiger vorkommen.
- Topic Models: Clustering-Methode, mit der Texte oder Textfragmente aufgrund ihrer Wortdistribution in Gruppen separiert werden, die sich durch eine ähnliche Wortverwendung auszeichnen.
- Netzwerkanalysen: Modellierung von Beziehungen zwischen Objekten (z.B. Wörtern) als Netzwerk.



Analysewort: Ausländer, Analysetyp 0

+	-1	-1	19509	**lebenden** hier legal	9	44%	von legal hier lebenden Ausländern
+	-1	-1	19509	lebenden hier rechtmäßig	11	54%	sollten alle rechtmäßig hier lebenden Ausländer auch arbeiten
+	-1	-1	19509	lebenden hier	545	61%	der|die hier [...] lebenden [...] Ausländer
+	-1	-1	19509	lebenden legal	51	52%	von legal [in Österreich] lebenden Ausländern
+	-1	-1	19509	lebenden rechtmäßig	33	51%	von rechtmäßig [in Deutschland] lebenden Ausländern allein abgeschoben
+	-1	-1	19509	lebenden	1887	66%	in hier|Deutschland lebenden [...] Ausländer
+	-2	-2	13649	**Integration** lebender	34	100%	die Integration hier|in Österreich lebender Ausländer
+	-2	-2	13649	Integration Aussiedlern	24	75%	die|zur Integration von Ausländern und Aussiedlern
+	-2	-2	13649	Integration	1933	62%	die Integration von Ausländern
+	-2	-2	13225	**Ausländerinnen** erleichterte	17	52%	erleichterte Einbürgerung junger Ausländerinnen und Ausländer
+	-2	-2	13225	Ausländerinnen Schweiz	45	60%	Ausländerinnen und Ausländer in|die der Schweiz
+	-2	-2	13225	Ausländerinnen Stimm	15	53%	Ausländerinnen und Ausländern ... das Stimm und Wahlrecht
+	-2	-2	13225	Ausländerinnen	950	67%	Ausländerinnen [und] Ausländer
+	-4	5	7413	**Deutschland** lebende	228	98%	in Deutschland [...] lebende [...] Ausländer
+	-4	5	7413	Deutschland geborene	123	57%	in|In Deutschland geborene Kinder von Ausländern die
+	-4	5	7413	Deutschland geborenen	86	82%	in Deutschland [...] geborenen [Kinder|Kindern von] Ausländern die ...
+	-4	5	7413	Deutschland	3688	42%	Ausländer [... in] Deutschland
+	-5	4	6904	**illegal** eingereiste	130	96%	illegal [...] eingereiste [...] Ausländer
+	-5	4	6904	illegal eingereisten	41	63%	von illegal [...] eingereisten [...] Ausländern
+	-5	4	6904	illegal beschäftigte	72	90%	illegal [...] beschäftigte Ausländer
+	-5	4	6904	illegal	1107	58%	illegal [in ...] Ausländer
+	-1	-1	6673	**lebende** hier legal	5	100%	legal hier lebende Ausländer
+	-1	-1	6673	lebende hier	153	96%	für hier [...] lebende [...] Ausländer
+	-1	-1	6673	lebende legal	36	100%	legal [in Österreich] lebende Ausländer
+	-1	-1	6673	lebende	744	97%	in hier|Deutschland lebende [...] Ausländer
+	-1	-1	6022	**viele** leben	143	62%	zu viele [...] Ausländer [...] leben
+	-1	-1	6022	viele lebten	29	62%	es|Deutschland lebten zu viele [...] Ausländer in
+	-1	-1	6022	viele wohnen	49	67%	in dem viele [...] Ausländer [...] wohnen
+	-1	-1	6022	viele	1867	84%	viele [...] Ausländer

Abbildung 2: Kookkurrenzprofil von "Ausländer" (Ausschnitt), CCDB

5.1 Kollokationsprofile vergleichen

Kollokationsprofile eines Lexems zeigen, mit welchen anderen Lexemen dieses häufiger, als wir es bei einer gleichmäßigen Verteilung der Lexeme im Korpus erwarten würden, auftritt. Das Profil ist damit eine Zusammenfassung aller Verwendungsweisen und zeigt die typischen Verwendungsweisen.

Belica entwickelte eine erweiterte Form von Kollokationsprofilen (hier genannt: Kookkurrenzprofil), bei denen nicht nur die signifikanten Kollokatoren zum Ausgangslexem, sondern auch sekundäre und tertiäre Kollokatoren sowie typische syntagmatische Muster angezeigt werden (CCDB: Belica 2001). Abbildung 2 zeigt die ersten Zeilen eines solchen Profils des Lexems ‚Ausländer' auf der Basis einer älteren Version des Deutschen Referenzkorpus DeReKo (Kupietz et al. 2010).

Das Kookkurrenzprofil zeigt deutlich, in welchen Kontexten im (zeitungslastigen) Korpus normalerweise von ‚Ausländern' die Rede ist. Die fett gedruckten Kollokatoren sind die primären Kollokatoren, um die sich dann ggf. noch weitere Kollokatoren gruppieren. Die weiteren Spalten informieren über die Position, an der der Kollokator vor (Minus-Werte) oder nach (Plus-Werte) dem Ausgangslexem auftritt, den statistischen Assoziationsgrad sowie die absoluten Frequenzen, mit denen die Kollokationen im Korpus erscheinen. In der letzten Spalte sind die typischen syntagmatischen Muster (mit Prozentangabe über den Anteil, den das Muster an allen Belegen einnimmt) aufgeführt (vgl. Bubenhofer/Scharloth 2013a für eine ausführlichere Erklärung zu den Kookkurrenzprofilen).

Über Kookkurrenzprofile des gleichen Lexems aber in unterschiedlichen Korpora, die unterschiedliche Diskurse repräsentieren, können diskursspezifische Verwendungsweisen ausgearbeitet werden. Eine weitere Möglichkeit ist der Vergleich zweier Kookkurrenzprofile unterschiedlicher Lexeme auf der gleichen Datenbasis: Wie unterscheidet sich die Verwendungsweise von ‚Ausländer' von

Abbildung 3: Vergleich der beiden Kookkurrenzprofile von "Ausländer" und "Flüchtling", CCDB

derer von ‚Flüchtling'? Mit der Funktion ‚contrast near-synonyms' ist es nicht nur möglich, Kookkurrenzprofile miteinander zu vergleichen, sondern das ganze semantische Feld zwischen den beiden Lexemen zu untersuchen.
Abbildung 3 zeigt diese Funktion der CCDB (Belica 2001) beim Kontrastieren der Profile von „Ausländer" und „Flüchtling". Die Idee ist Folgende: „Ausländer" und „Flüchtling" haben je ein spezifisches Kollokationsprofil. Zusätzlich gibt es aber viele weitere Lexeme, die ähnliche Profile haben wie die beiden Ausgangslexeme und dabei mehr oder weniger ähnlich sind wie das eine oder andere Lexem. Diese ähnlichen Lexeme werden nun nach Grad der Ähnlichkeit mit den Ausgangslexemen angeordnet, wobei zusätzlich ein Clustering stattfindet: Lexeme, deren Profile besonders ähnlich sind und sich gleichzeitig von den anderen unterscheiden, werden zu einer Gruppe zusammengefasst. Jedes Feld in Abbildung 3 enthält einen solchen Cluster von Lexemen mit besonders ähnlichen Kollokationsprofilen. Die farbliche Codierung zeigt zudem an, wie ähnlich die Lexeme im Vergleich zu „Ausländer" (hell, im Original gelb) oder „Flüchtling" (dunkel, im Original rot) sind.

Nun wird sichtbar, dass „Flüchtling" in ähnlichen Kontexten erscheint wie eine Reihe von Herkunftsbezeichnungen von typischen Emigrationsländern („Bosniake", „Serbe", „Kroate", „Ruander") oder Kriegsterminologie („Zivilbevölkerung", „Offensive", „Regierungstruppe" etc.), die Fluchtgründe benennt. Sehr ähnlich wie „Ausländer" sind jedoch Lexeme wie „Menschenhandel", „Ausländerhass", „ausländerfeindlich", „rassistisch" etc., also Phänomene, die eintreten können, wenn Migrantinnen und Migranten im Inland thematisiert werden. Dazwischen stehen z.B. mit einer Nähe zu „Flüchtling" Institutions- oder Menschenbezeichnungen, bei denen es um Flüchtlinge im Inland geht: „Neuankömmling", „Waisenkind", „Flüchtlingsrat", „Abschiebehäftling" usw. Ebenfalls zwischen „Ausländer" und „Flüchtling", jedoch näher bei ersterem, sind Bezeichnungen für gesellschaftlich diskriminierte Menschen wie „Schwule", „Andersdenkende", „Homosexuelle", „Drogensüchtige", „Kriminelle".

Es gäbe weitere Beobachtungen zu machen, doch bereits diese kurze Analyse zeigt wichtige Aspekte des Migrationsdiskurses (allerdings in einer bereits historischen Perspektive, da die zugrundeliegenden Daten nicht aktuell sind): „Flüchtlinge" heißen Migrantinnen bzw. Migranten dann, wenn sie im Kontext ihrer Fluchtgründe benannt werden. Erreichen sie das Inland, werden sie zu Ausländern, wobei dann nicht mehr die ursprünglichen Fluchtgründe bedeutend sind, sondern der Umgang mit ihnen im Inland.

Die (diskurs-)linguistische Bedeutung dieser Methode ist bemerkenswert: Wir finden damit typische Verwendungsweisen eines bestimmten Wortes und wiederum Wörter, die ähnlich verwendet werden wie dieses, also Aussagemuster, die sich ähneln, obwohl deren Kerne auf den ersten Blick nichts miteinander zu tun haben („Ausländer" und „Homosexueller"). Es offenbaren sich Muster des Sagbaren in Diskursen. Daneben erhält man datengeleitet eine Übersicht über

involvierte Akteure, Vorgänge oder Zustände, die tatsächlich inhaltlich etwas mit den verglichenen Ausgangslexemen zu tun haben oder aber einem ähnlichen Dispositiv entspringen.

5.2 Sprachgebrauchsmuster

Sprachgebrauchsmuster können als Indikatoren für verschiedene Aspekte eines Diskurses gelesen werden (Bubenhofer 2009). Sie können korpuslinguistisch operationalisiert werden als Ketten von sprachlichen Einheiten wie Wortformen, Grundformen oder Wortartklassen, die typisch für ein bestimmtes Untersuchungskorpus (Diskurs X in Zeitung A) im Vergleich zu einem Referenzkorpus (alle Themen in Zeitung A) sind. Im einfachsten Fall werden dazu alle Wortformen-n-Gramme im Untersuchungs- und Referenzkorpus extrahiert und deren Frequenzen in den beiden Korpora verglichen. N-Gramme, die im einen Korpus im Vergleich zum anderen überzufällig häufig auftreten, sind demnach statistisch auffällig und lohnenswert, um sie als diskursspezifisch zu interpretieren.

Anhand des Text+Berg-Korpus (Bubenhofer et al. 2013b) konnten wir mit dieser Methode zeigen, wie sich das Sprechen über Berge in Bergsteigerberichten des Schweizer Alpenclubs in 150 Jahren verändert hat (Bubenhofer/Scharloth 2011, 2013b; Bubenhofer/Schröter 2012). Auffallend für die 1880er- und 1890er-Jahre sind Muster wie „[Adjektiv] Stunde [Adjektiv] [Nomen]“ mit Realisierungen wie „halben Stunde angenehmer Steigung“, „halben Stunde weiteren Weges“ oder das Muster „[Präposition] [Artikel] Nähe [Artikel] [Nomen]“ mit Realisierungen wie „in der Nähe des Gipfels/der Grenze/des Muttensees“ etc. Die meisten Muster spiegeln einen logbuchartigen, sachlichen Stil wider, mit dem die Landschaft genau vermessen wird.

In den 1930er- und 40er-Jahren sind Muster vorherrschend, die einen narrativen Stil vermuten lassen: „dann [Vollverb] [Artikel] [Nomen/Adjektiv]“ mit Realisierungen wie „dann kündet die Gipfelglocke“, „dann geschieht das Wunder“ etc. oder „[Adverb] [Vollverb] [Artikel] [Adjektiv] [Nomen]“ mit Realisierungen wie „draussen erwachte ein neuer Tag“, „dann kam ein trüber Tag“, „nun naht das schwierigste Stück“.

Immer emotionaler und subjektiver wird diese Erzählweise in den 1960er- und 70er-Jahren, wo auch vermehrt Muster mit den Personalpronomen in erster Person Singular und Plural auftreten oder Ausrufe wie „… [Adjektiv] [Nomen] [Ausrufezeichen]“ – „War das ein wertvoller Fund!“, „Eine wunderbare Welt!“, „Welch andere Welt!“. Im Zusammenhang mit anderen korpuslinguistischen Erkenntnissen wird deutlich, dass der Diskurs des Sprechens über Berge zu einem Sprechen über sich selber in den Bergen wird.

5.3 Semantische Lesarten diachron

Kollokationsprofile spiegeln typische Verwendungsweisen von Wörtern (siehe oben) wider. Die unterschiedlichen Verwendungsweisen müssen aber interpretativ aus dem Profil abgeleitet werden. Um solche Verwendungsweisen datengeleiteter aus einem Korpus zu berechnen, eignet sich die Anwendung eines Topic Modelling-Verfahrens.

Unter ‚Topic Models' werden verschiedene Ansätze zusammengefasst, die darauf zielen, Texte nach Themen zu clustern (Anthes 2010; Graham et al. 2012). Dabei sollen die Themen jedoch nicht vorgegeben, sondern aus den Daten abgeleitet werden. Vereinfacht gesagt werden Texte aufgrund der Wortverteilung, gemessen an einer bestimmten Wahrscheinlichkeitsverteilung, in Klassen aufgeteilt und die dafür charakterisierenden Wörter genannt. Der Algorithmus versucht die bestmögliche Klassifizierung zu finden, die aufgrund des Wortmaterials die Texte in möglichst homogene Gruppen aufteilt.

Bei Rohrdantz et al. (2012) findet sich ein Ansatz, diese Methode zu verwenden, um semantische Veränderungen von Lexemen datengeleitet zu berechnen. Anstelle von Texten werden Belege des gesuchten Lexems klassifiziert und die Verteilung der unterschiedlichen Verwendungsweisen diachron abgebildet.

So können beispielsweise die semantischen Facetten von „Terror" herausgearbeitet werden (Bubenhofer/Scharloth 2015): „Al-Qaida-Terror", „RAF-Terror", „Brigade Rosse-Terror" und weitere mehr. Während diese Lesarten einen starken historischen Bezug haben und wenig überraschend sind, sind Analysen abstrakterer Konzepte interessanter. Betrachten wir die semantischen Facetten von „Freiheit".

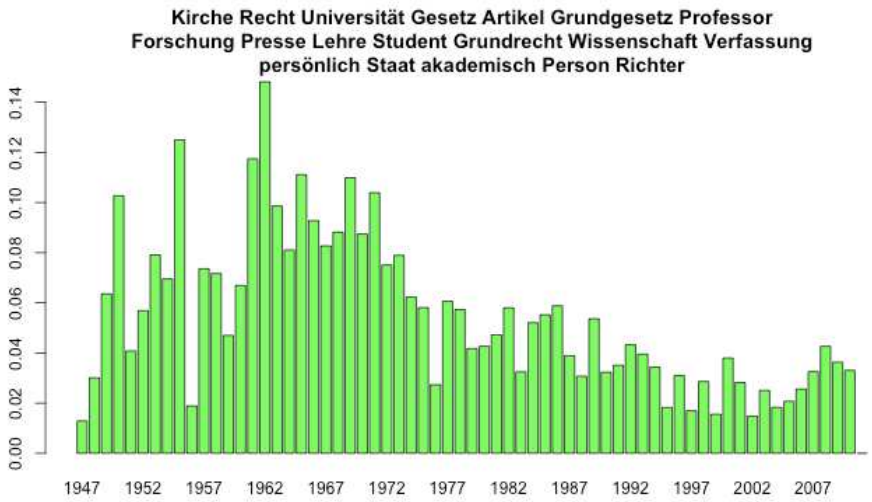
Die Datengrundlage ist ein Korpus aller Artikel des Magazins ‚Der Spiegel' von 1947 bis 2010 (237.620.381 Tokens, 307.111 Texte). Daraus werden alle Belege für das Lexem „Freiheit" extrahiert, wobei jeweils ein Kontext von 25 Wörtern davor und danach berücksichtigt wird. Die 24.291 Belege, in lemmatisierter Form, werden nun einer LDA-Klassifikation (Blei et al. 2003) unterzogen. Da dem Clusteringalgorithmus eine Zielgröße der Anzahl der zu findenden Cluster übergeben werden muss, ist es sinnvoll, verschiedene Zielgrößen zu testen, um eine optimale Anzahl zu finden. Im Beispiel erwiesen sich zehn unterschiedliche Cluster als sinnvoll. Der Algorithmus kommt zu folgenden Clustern, charakterisiert durch die jeweiligen ‚Keywords':

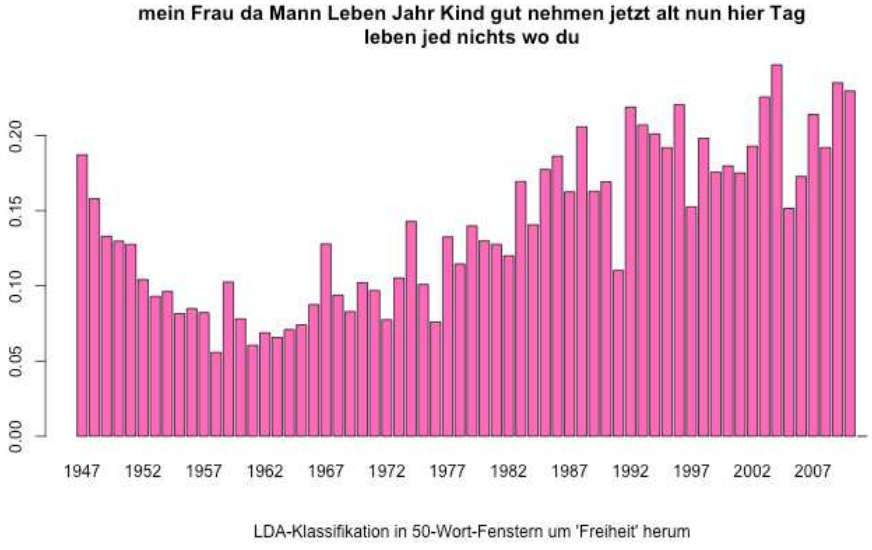
#	Prob.	Keywords	Bezeichnung (manuell)
0	0,13	Kirche Recht Universität Gesetz Artikel Grundgesetz Professor Forschung Presse Lehre Student Grundrecht Wissenschaft	Grundrechte / Menschenrechte

		Verfassung persönlich Staat akademisch Person Richter	
1	0,27	Land Jahr Volk Präsident Krieg Polen kämpfen amerikanisch Regierung Sowjet Demokratie Amerika Kampf Welt politisch US Unabhängigkeit Tag USA	Erkämpfte Freiheit
2	0,3	mein Frau da Mann Leben Jahr Kind gut nehmen jetzt alt nun hier Tag leben jed nichts wo du	Persönliche Freiheit
3	0,73	Freiheit Mensch jed politisch Demokratie Staat ohne Recht frei sondern Deutschland denn ganz eigen weil gut Frage Gesellschaft Bürger	Freiheit und Demokratie
4	0,03	Mark Weg finanziell Bodo Schäfer Hans Campus Seite Verlag Euro Macht _(Leben NUMMER mein Hoffmann Bertelsmann Neger Campe	(keine sinnvolle Zuordnung)
5	0,08	Uhr Film Kunst künstlerisch ZDF ARD Autor Buch Regisseur Künstler TV - ord Theater Stück Schriftsteller Geschichte spielen Roman	(keine sinnvolle Zuordnung)
6	0,13	Jahr Gefängnis Monat entlassen Haft Tag drei verurteilen Woche Mann Gefangene Gericht Stunde letzt Häftling sitzen nun ins Arzt	Physische Freiheit
7	0,15	SPD CDU Sozialismus statt Partei FDP Strauß jung CSU Zeitung Kohl Deutschland Bonner Berlin Berliner Wahlkampf - Wissenschaft Woche	Parteipolitische Freiheit
8	0,1	Berlin West Million Ost New Berliner Prozent Jahr Mark York rund DDR Preis Westen Stadt Symbol amerikanisch Dollar Auto	BRD West/Ost-Freiheit
9	0,14	Freiheit Mensch Gleichheit Reich Gott Brüderlichkeit Revolution sondern Idee menschlich französisch Welt Jahrhundert Geschichte Ideal Marx Gerechtigkeit jen bürgerlich	Menschenrechte historisch

Die Wahrscheinlichkeit (‚Prob.‘), mit der die Belege den Clustern zugeordnet werden können, ist nicht überall befriedigend. Insbesondere die Cluster 4 und 5 gehen wahrscheinlich auf Buch- und Filmanzeigen zurück, die auch nicht sehr oft

im Korpus vorkommen. Für die anderen Cluster ist es aber aufgrund der Keywords möglich, Namen zu vergeben. Natürlich kann ein Blick in die Belege und die Zuordnungen bei der Interpretation helfen: Für jeden Beleg ist angegeben, mit welchen Wahrscheinlichkeiten die zehn Cluster zugeordnet worden sind. In einem weiteren Schritt kann die Verteilung der Häufigkeiten, mit denen die Cluster den Belegen zugeordnet worden sind, auf eine zeitliche Achse abgetragen werden (Abbildung 4).


Kirche Recht Universität Gesetz Artikel Grundgesetz Professor
Forschung Presse Lehre Student Grundrecht Wissenschaft Verfassung
persönlich Staat akademisch Person Richter
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1947
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007
LDA-Klassifikation in 50-Wort-Fenstern um 'Freiheit' herum


mein Frau da Mann Leben Jahr Kind gut nehmen jetzt alt nun hier Tag
leben jed nichts wo du
0.00
0.05
0.10
0.15
0.20
1947
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007
LDA-Klassifikation in 50-Wort-Fenstern um 'Freiheit' herum

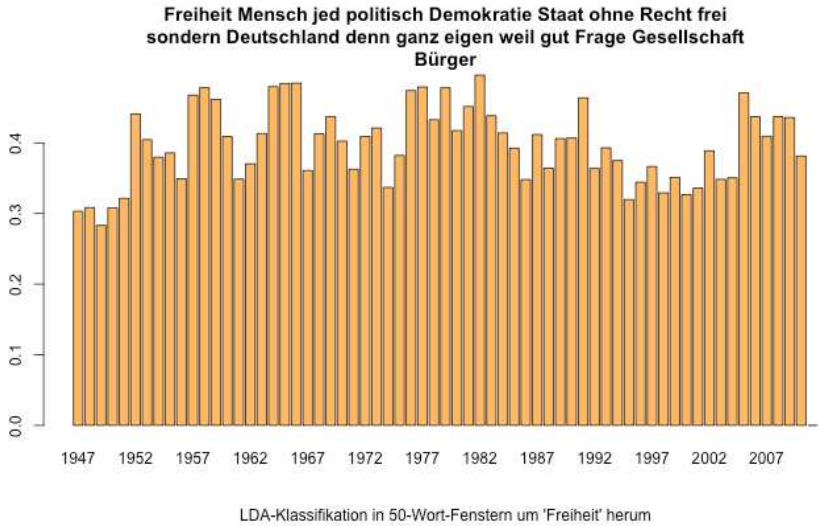


Abbildung 4a-c): Distribution der Cluster auf die zeitliche Achse (Cluster 0, 2 und 3)

Die Lesart ‚Grundrechte/Menschenrechte' nimmt bis in die 60er-Jahre zu, geht dann aber kontinuierlich zurück. Die Lesart ‚persönliche Freiheit' nimmt nach einem Tiefpunkt am Anfang der 60er-Jahre kontinuierlich zu und die Lesart ‚Freiheit und Demokratie' bleibt über die ganze Periode mehr oder weniger stabil, allerdings mit regelmäßigen Ausschlägen (die Debatte um die Volkszählung in den 1980er-Jahren in Deutschland könnte für den Ausschlag in dieser Zeit verantwortlich sein).

Das Topic Modelling-Verfahren hat nicht nur datengeleitet zu möglichen semantischen Facetten von „Freiheit" geführt, sondern macht es auch möglich, die 24.000 Belege zu klassifizieren und so Veränderungen in der Verwendungsweise festzustellen und somit ggf. eben auch Hinweise für diskursive Veränderungen zu haben.

5.4 Netzwerke

Es ist inzwischen üblich geworden, ganz unterschiedliche Gegenstände als Netzwerk zu konzipieren. Im Grimm'schen Wörterbuch wird ‚Netzwerk' umschrieben als „*etwas netzartiges*: alle fächlein oder bläslein der läpplein in den lungen werden mit einem sehr subtilen netzwerke .. umgeben. ZEDLER 23, 2021" (DWB, „Netzwerk"). Besonders produktiv wurde der Netzwerkbegriff durch dessen Abstrahierung durch die Anwendung auf soziale Gefüge in den 1960er-Jahren

(Mitchell 1969), sodass sich inzwischen eine eigenständige Netzwerkforschung entwickelt hat (Newman 2010; Stegbauer 2010). Im Kontext der Kulturgeschichte zeigten etwa Schich et al. (2014), wie aus trivialen Daten, nämlich den Geburtsorten/-daten und Sterbeorten/-daten von über 120.000 ‚bekannten Persönlichkeiten' ein interpretierbares Netzwerk entsteht: Werden die Geburts- und Sterbeorte als auf einer Karte georeferenzierte Knoten und die Bewegung der Personen als Kanten dazwischen visualisiert, wird im diachronen Vergleich sichtbar, welche Orte sich als kulturelle Zentren etablieren konnten oder diesen Status wieder verloren. Die quantitative Auswertung der Daten und insbesondere die Visualisierung als Netzwerk ermöglichen eine neue Sicht auf die Daten, die vorher nicht möglich war.

Auch Kollokationen lassen sich als Netzwerk auffassen, wobei die kollokierenden Wörter als Knoten mit einer Kantenverbindung, die ggf. die Assoziationsstärke visualisiert (etwa durch Dicke oder Länge). Abbildung 5 zeigt ein Kollokationsnetz auf der Basis von 1962 Texten (7.562.273 Tokens) aus dem Text+Berg-Korpus (Bubenhofer et al. 2013b), also von Bergsteigerberichten. Daraus wurden für alle Lexeme, die im Vergleich zu einem Referenzkorpus signifikant sind, Kollokatoren berechnet und als Netzwerk visualisiert. In Abbildung 7 ist ein Detail aus der Übersichtsdarstellung abgebildet.

Bevor ein kurzer Blick auf die Inhalte des Netzwerkes geworfen werden soll, lohnt es sich, die Konstruktion des Netzwerkes als solches vor Augen zu führen. Die zugrundeliegenden (berechneten) Kollokationsdaten haben vereinfacht folgendes Format:

Westwind → heftig
Gewitter → heftig
los+brechen → Blitz
los+brechen → Gewitter
Blitz → Gewitter
flammen → Blitz
grell → zucken
grell → Blitz
heftig → Kälte
…

Es ist unmöglich, durch Lesen solcher langer Listen einen Eindruck über das Ensemble der Verknüpfungen zu erhalten. Erst durch die Visualisierung als Netz kann sich ein Bild ergeben (Pfeffer 2010; für einen historischen Blick auf Visualisierungen von Netzwerken: Kruja et al. 2002). Allerdings stellt sich die Frage, wie die Knoten (Wörter) im Raum – und damit ist meist eine zweidimensionale Fläche gemeint – angeordnet werden sollen. Dafür können unterschiedliche Layoutalgorithmen verwendet werden (vgl. Abbildung 6). Bei komplexen Netzwerken ist dies keine triviale Angelegenheit, da für jeden Knoten vielfältige Bedingungen gelten sollen, die sich widersprechen können (z.B. sowohl eine

Verbindung zu einem weiteren Knoten ganz links als auch ganz rechts in der Fläche, gleichzeitig aber nicht zu nahe an einem dritten Knoten). Oft wird einem physikalischen Prinzip folgend ein Optimum der Positionierung der Knoten zu erreichen versucht. Ein häufiges Modell in der Gruppe der ‚force-directed' Layoutmodi nimmt zwei grundsätzlich wirkende Kräfte an: Einerseits werden die Knoten, als ob sie an Springfedern angemacht wären, ins Zentrum der Fläche gezogen, gleichzeitig stoßen sich Knoten aber gegenseitig ab, als ob sie elektrisch gleich geladen wären. Lässt man die Berechnung nach diesem Prinzip mehrfach über die Knoten iterieren, ergibt sich mit der Zeit eine optimale Darstellung. Es existieren verschiedene Spielarten dieses Prinzips, die z.B. auf schnelle Berechenbarkeit optimiert sind (Chen et al. 2008, 109; Pfeffer 2010).

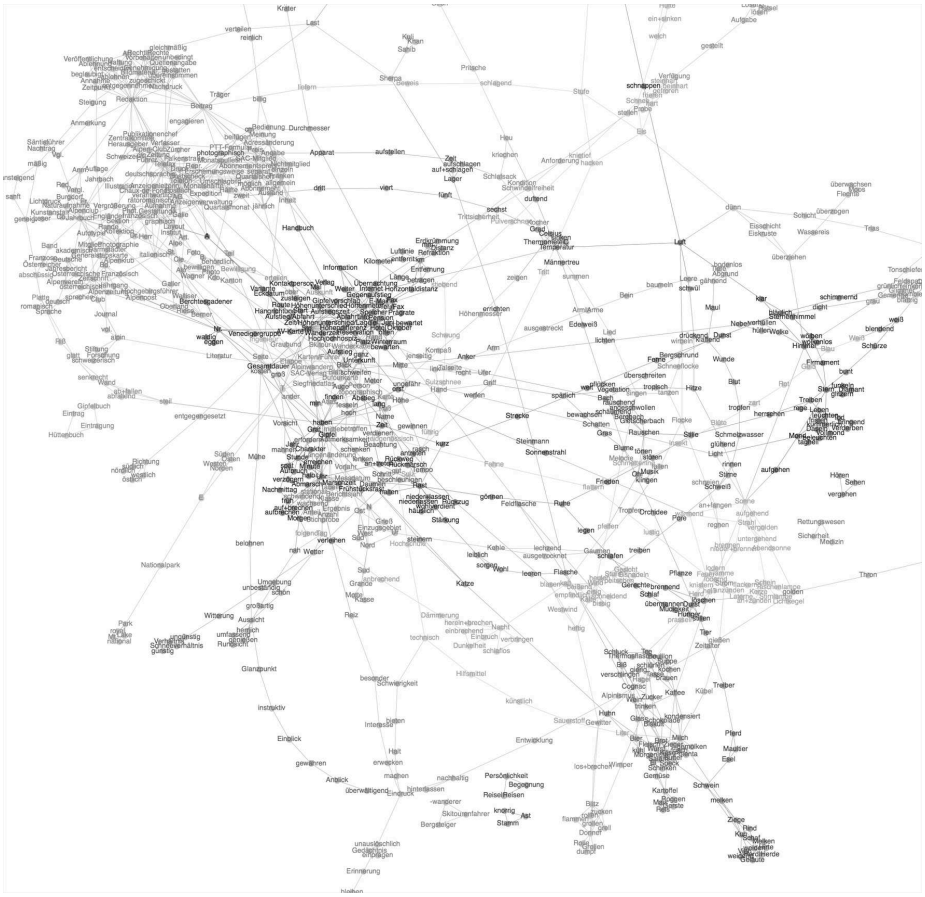


Abbildung 5: Die typische Bergsteigergeschichte: Netz von Kollokationen im Text+Berg-Korpus (vgl. https://www.bubenhofer.com/sprechtakel/2013/02/21/die-typische-bergtour/ für eine interaktive Version)

Schaut man sich nun das nach einem force-directed-Algorithmus (hier: Force Atlas) gelayoutete Kollokationennetzwerk genauer an, ergeben sich alleine durch die Anordnung der Knoten Cluster von Themen, die plausibel gedeutet werden können. Vgl. etwa in Abbildung 7 die Cluster zum Wetter und zum Essen, die offenbar Konstanten in den Bergsteigerberichten sind. Man sieht aber auch Knoten, die für mehrere Cluster bedeutend sind, etwa „Hagel“, der sowohl über „prasselt“ eine Verbindung zum „Feuer“-Cluster hat („Feuer prasselt“), als auch über „Gewitter“ zum Wetter-Cluster. Darstellungsbedingt kommt „Hagel“ aber über den Essen/Trinken-Cluster zu liegen; würde man für das Layout eine

dreidimensionale Darstellung wählen, würde der Layoutalgorithmus eine bessere Lösung finden können (vgl. zu diesem Problem der Dimensionsreduktion Abschnitt **Fehler! Verweisquelle konnte nicht gefunden werden.**).

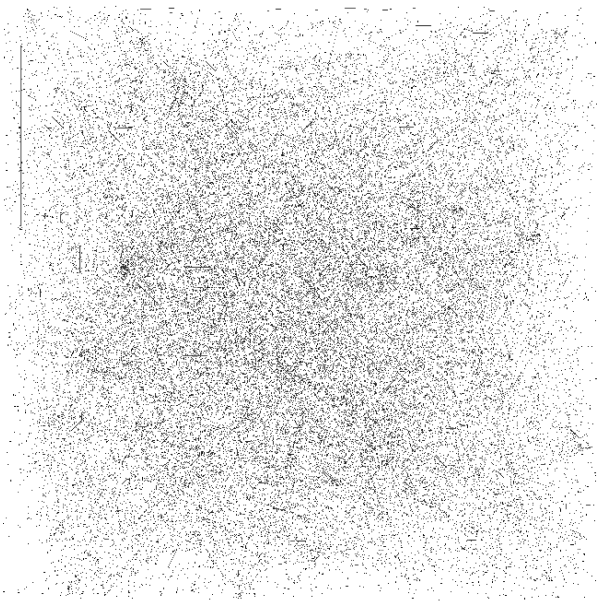
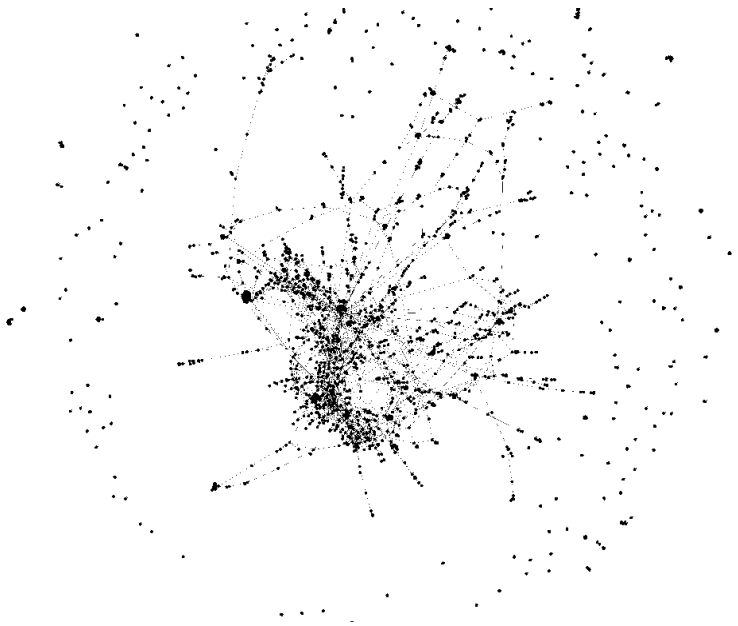
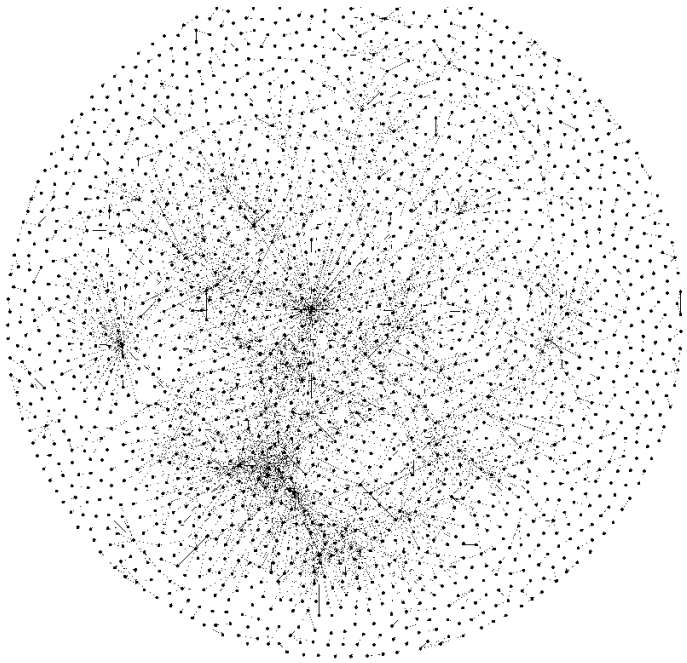


Abbildung 6: Verschiedene Netzwerklayouts der gleichen Daten: 1) Zufällige Anordnung, 2) Force Atlas, 3) Fruchterman Reingold, 4) Yifan Hu





entwickelt, um Aussagen über das Netzwerk machen zu können. So In

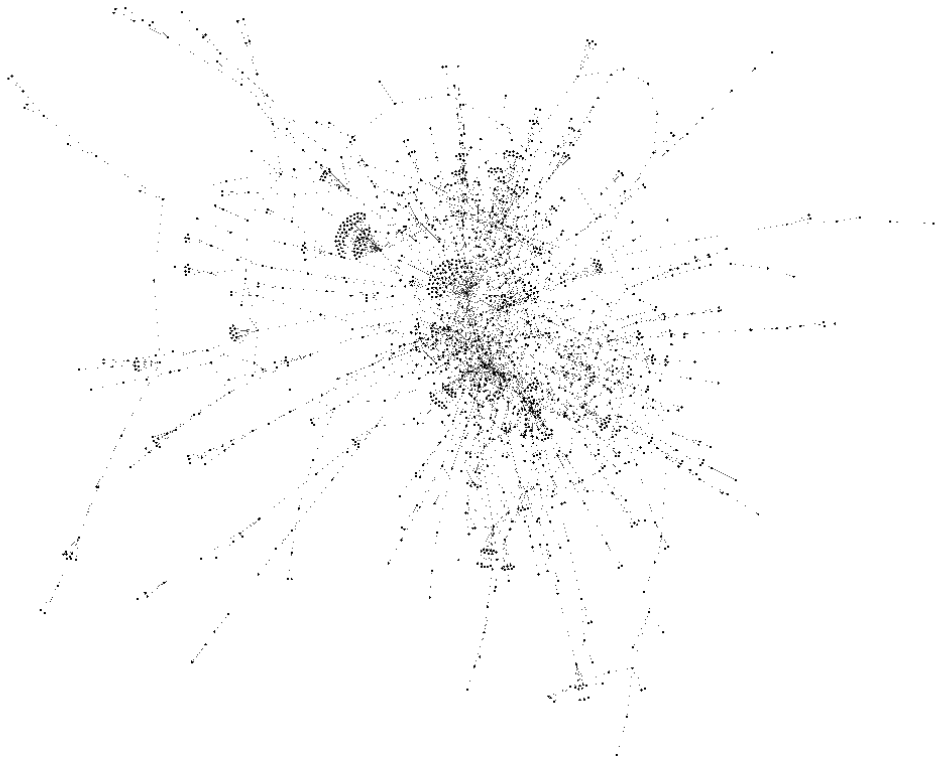


Abbildung 6a-d: Verschiedene Netzwerklayouts der gleichen Daten: 1) Zufällige Anordnung, 2) Force Atlas, 3) Fruchterman Reingold, 4) Yifan Hu

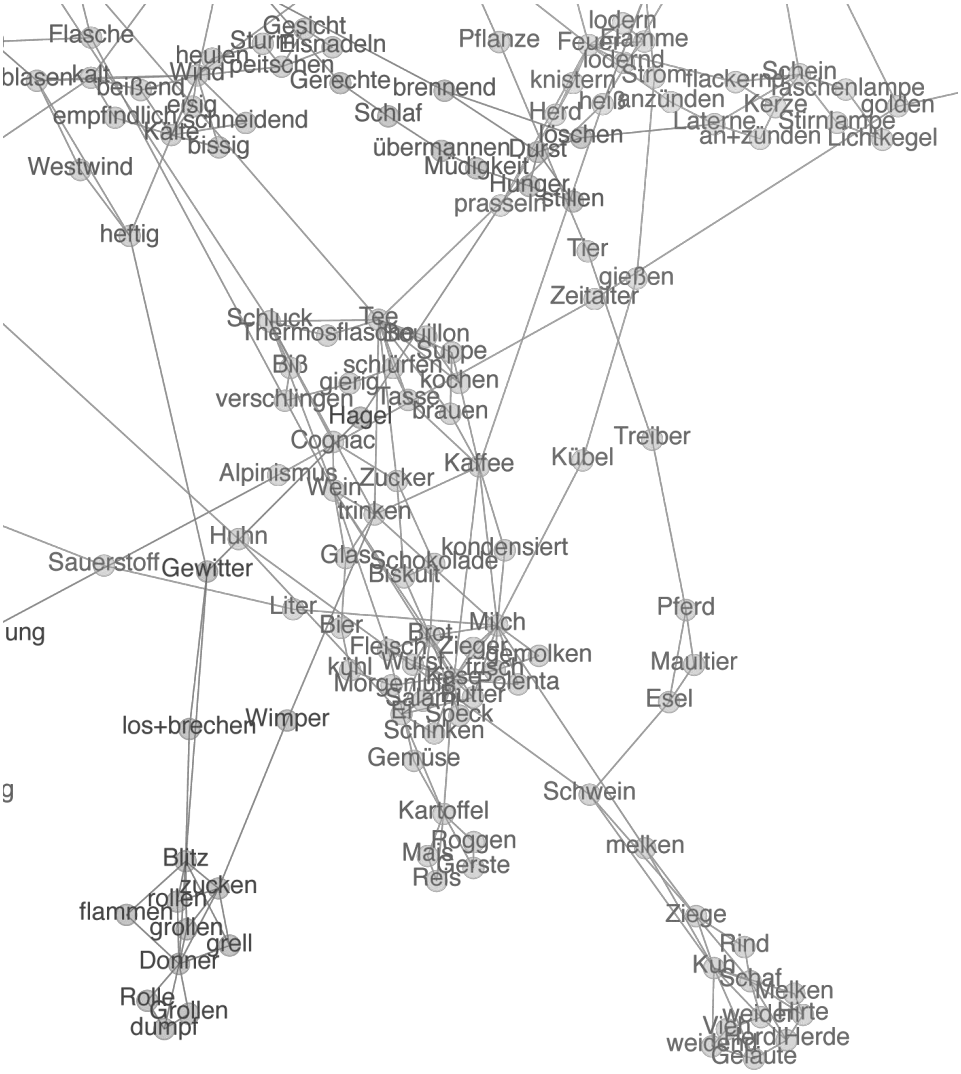


Abbildung 7: Ausschnitt aus dem Kollokationsgraph der Bergsteigerberichte

In der Netzwerkforschung wurden verschiedene statistische Maße entwickelt, um Aussagen über das Netzwerk machen zu können. So gibt es unterschiedliche Zentralitätsmaße, um zu messen, wie ‚wichtig' ein Knoten im Netzwerk ist (Newman 2010, 168). Die ‚betweenness centrality' drückt beispielsweise aus, wie oft ein Knoten durchlaufen wird, wenn man alle möglichen Verbindungen zwischen allen Knoten im Netz in Betracht zieht und jeweils den kürzesten Weg wählt (Newman 2010, 185). Im Beispiel oben wurde die Louvain-Methode

(Blondel et al. 2008) angewandt, um ‚Communities' von Knoten zu entdecken, die viele Verbindungen untereinander, aber wenige mit anderen Knoten aufweisen.

Es ist naheliegend, Methoden der Netzwerkforschung für diskurslinguistische Zwecke zu nutzen – und zwar sowohl indem 1) Akteure oder Artefakte (bzw. ihre Eigenschaften) als Netzwerk modelliert werden (analog dem ersten Beispiel oben von Schich et al.), als auch 2) indem sprachliche Eigenschaften des zugrundeliegenden Textkorpus als Netzwerk aufgefasst werden. Letzteres ist eine primär diskurs*linguistische* Herangehensweise, während ersteres ganz generell für diskursanalytische Fragestellungen auch in anderen sozial- und geisteswissenschaftlichen Disziplinen eine Rolle spielt.

6 Fazit

Die Korpuslinguistik kann für die Diskursanalyse eine Hilfswissenschaft sein, um bestimmte Thesen oder Phänomene in großen Datenmengen effizient zu untersuchen. Sie kann aber, das hoffe ich gezeigt zu haben, auch viel mehr sein: Der Schlüssel zu einem neuen Verständnis, wie wir in den Geisteswissenschaften mit Daten umgehen. Daraus ergeben sich Möglichkeiten für neue Fragestellungen, kombiniert mit einer äußerst wichtigen Diskussion und Reflexion über das Zusammenspiel von Methoden und theoretischen Annahmen. Bereits die Art und Weise, wie Daten repräsentiert werden – numerisch in einem Vektorraum statt als Texte in einer Datenbank – hat einen Einfluss darauf, unter welchen theoretischen Prämissen die Daten analysiert werden können. Häufig benutzte Methoden wie das Topic Modelling oder die Repräsentation von Objekten als Netzwerk basieren auf Algorithmen, die in den Geisteswissenschaften nur zu gern als Black Box verwendet werden. Die Auseinandersetzung mit den zugrundeliegenden Algorithmen gehört jedoch genauso zu den theoretischen Prämissen wie die Entscheidung, eine Diskurslinguistik eher nach der ‚Düsseldorfer Schule' oder als ‚Critical Discourse Analysis' zu betreiben (Spitzmüller/Warnke 2011, 81). Visualisierungsmethoden zur Exploration von Daten beeinflussen ebenfalls markant die methodischen Zugänge, wobei deren Status im Forschungsprozess, gerade wenn man es mit sprachlichen Daten zu tun hat, noch nicht hinreichend reflektiert ist (Bubenhofer/Scharloth 2015, 16; siehe für den Vorschlag einer New Visual Hermeneutics: Kath et al. 2015).

Wir haben unlängst vier Desiderate formuliert, die im Rahmen einer sozial- und kulturwissenschaftlich interessierten maschinellen Textanalyse verfolgt werden sollten (Bubenhofer/Scharloth 2015), die ich hier wiederholen und um einen Aspekt ergänzen möchte:

- *Die maschinelle Textanalyse braucht einen integrierten Textbegriff:* Die meisten Methoden des Data-Minings und der Computerlinguistik gehen von sogenannten ‚Bag of Words'-Ansätzen aus: Ein Text (oder eine kleinere Einheit) besteht aus einer Menge von Wörtern, die jedoch ihrer Sequenzialität entrissen werden. Bei der oben dargestellten Methode des Topic Modellings spielt die Position des Wortes im Text oder Textausschnitt für die Bestimmung der Clusterzugehörigkeit keine Rolle. Text – und auch Diskurs – wird also meistens nicht als komplexes Gewebe operationalisiert. Das liegt nicht an technischen Beschränkungen, sondern daran, dass sich die Geisteswissenschaften bislang zu wenig für maschinelle Methoden interessierten und ihre Sicht einfließen ließen. Ein Versuch, Narrative in Geschichten vom ‚Ersten Mal' datengeleitet zu berechnen, liegt vor und zeigt das Potenzial, über die ‚Wortsäcke' hinwegzukommen (Bubenhofer et al. 2013a).
- *Die maschinelle Textanalyse braucht valide Modelle:* Methoden des Data-Minings müssen *funktionieren*, nicht aber unbedingt *erklären* können. Für die maschinelle Autorschaftsattribution mag ein Ansatz, der ein Modell auf der Basis von Buchstaben-n-Grammen verwendet, funktionieren, um Texte mit unbekannter Autorschaft einem Autor bzw. einer Autorin zuzuweisen. Das Modell hat aber aus linguistischer Sicht keinen erklärenden Wert, da Buchstaben-n-Gramme keine linguistischen oder literaturwissenschaftlichen Stilbegriffe operationalisieren. Für die Sozial- und Kulturwissenschaften sind demnach Modelle wichtig, die hinsichtlich des zu Operationalisierenden valide sind (Kath et al. 2015; Lemke/Stulpe 2015) und die White-box- statt Black-box-Algorithmen verwenden (Rieder/Röhle 2012).
- *Die maschinelle Textanalyse braucht neue Methoden der Visualisierung:* Auch Visualisierungen sind nicht nur ein Hilfsmittel, um Ergebnisse übersichtlich darzustellen, sondern – wie das Beispiel zu den Netzwerken oben gezeigt hat – ein hermeneutisches Mittel zur Datenanalyse. Schon länger firmieren unter Labeln wie ‚Visual Analytics' oder ‚Scientic Visualization' Paradigmen, die im Data-Mining erfolgreich eingesetzt werden (Chen et al. 2008; Keim et al. 2010). Obwohl die Linguistik eine lange Tradition von Visualisierungstechniken als exploratives Mittel aufweist (Dialektkarten, Gesprächstranskriptionen, Key Word in Context-Darstellungen etc.), steckt die Reflexion darüber, welchen Effekt diese Visualisierungen auf die Sprachdaten haben und wie sie im Forschungsprozess eingebettet sind (Stichwort „Macht der visuellen Evidenz": Rieder/Röhle 2012), in den Anfängen. Die diagrammatische Forschung dazu zeitigt interessante Überlegungen, die intensiviert werden müssen (Bauer/Ernst 2010; Krämer 2009; Steinseifer 2013; Stjernfelt 2007). Zudem ergeben sich mit digitalen Daten neue Möglichkeiten von Visualisierungsformen, die erkundet werden sollten.

- *Die maschinelle Textanalyse braucht eine Forschungsethik:* Digitalisierung sind nicht nur Zahlen, Daten, Informationen und Algorithmen, sondern der Umgang damit ist Praxis. Diese Praxis gilt es zu hinterfragen, da sie eine Reihe von ethischen Problemen berührt. Darunter fallen die forschungsethischen Grundprinzipien, etwa wenn individuelle Spuren in großen Datenmengen nachgezeichnet werden können, oder Gleichbehandlung, wenn der Zugang zu Daten und Analysetools beschränkt ist oder bestimmte Personengruppen von der Datenanalyse ausgeschlossen sind, da ihre Daten weniger gut greifbar sind. Zusätzlich wird der Ausgleich zwischen Urheber-/Verwertungsrechten und informationsethischen Grundprinzipien in Frage gestellt – nur ein Beispiel: Was ist höher zu gewichten, das Gemeininteresse an Informationen oder die kommerzielle Verwertung derselben?

- *Die maschinelle Textanalyse braucht eine Reflexion über Coding Cultures:* Programmiersprachen, die verwendet werden, um Algorithmen zu implementieren und daraus Software zu bauen, werden gemeinhin als neutrales Werkzeug angesehen. Die Entscheidung über die Wahl der Programmiersprache als höchstens ingenieurtechnisch interessante Trivialität. Auf den zweiten Blick ist aber klar, dass Programmiererinnen und Programmierer unterschiedlicher Programmiersprachen „different cultures, different tribal folklores, that they use to organize their working life“ (Ford 2015) haben. Software ist „a theoretical category […] still invisible to most academics, artists, and cultural professionals interested in IT and its cultural and social effects“ (Manovich 2014) – und ich würde weitergehen: Die Praxis der Softwareerstellung ist es ebenso, da die Praxis des Programmierens eine zutiefst kulturelle Praxis ist. Im Bereich der Visualisierung haben technologische Neuerungen (HTML5, SVG) zu einer völlig neuen Praxis der Visualisierung geführt, wie an sog. Bibliotheken für JavaScript, eine populäre Scriptsprache für Webanwendungen, gezeigt werden kann. D3.js oder P5.js sind Beispiele für solche Bibliotheken, deren technische Grundlagen und Möglichkeiten ähnlich sind, mit denen jedoch sehr unterschiedliche Kulturen verknüpft sind. In einem interaktiven Präsentationsvideo von P5.js (http://hello.p5js.org/) gibt die Hauptentwicklerin Lauren McCarthy dieses OpenSource-Projekts den unkomplizierten, undogmatischen Stil vor, in dem mit der Sprache programmiert werden soll. Dies kontrastiert bereits mit der ernsteren Selbstpräsentation von D3 (http://www.d3js.org) und größtmöglich etwa mit Auftritten des ehemaligen CEO von Microsoft, Steve Ballmer, der an Entwicklerkonferenzen zur Befeuerung „Developers“ bis zur Heiserkeit schreit. Dem Programm sieht man an, welcher Programmierkultur es entspringt und die Programmierkultur befördert einen bestimmten (wissenschaftlichen) Denkstil (nach Ludwik Fleck, 1980).

Der Diskurslinguistik ist zuzutrauen, diese metareflexiven Aufgaben angehen zu können.

Literatur

Anthes, Gary (2010): Topic Models Vs. Unstructured Data. In: Communications of the ACM 53, 16–18.

Bauer, Matthias/Christoph Ernst (2010): Diagrammatik / Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld. Bielefeld.

Belica, Cyril (2001): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Abgerufen am 4.3.2014 von http://corpora.ids-mannheim.de.

Blei, David M./Andrew Y. Ng/Michael I. Jordan (2003): Latent dirichlet allocation. In: Journal of Machine Learning Research 3, 993–1022.

Blondel, Vincent D./Jean-Loup Guillaume/Renaud Lambiotte u.a. (2008): (2008): Fast unfolding of communities in large networks. In: Journal of Statistical Mechanics: Theory and Experiment 2008 10, P10008, doi: 10.1088/1742-5468/2008/10/P10008.

Bray, Tim/Jean Paoli/C. Michael Sperberg-McQueen (o. J.): Extensible Markup Language (XML) 1.0. W3C Recommendation.

Bubenhofer, Noah (2006-2016): Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. Abgerufen am 13.4.2016 von http://www.bubenhofer.com/korpuslinguistik/.

Bubenhofer, Noah (2014): Geokollokationen – Diskurse zu Orten: Visuelle Korpusanalyse. In: Sondernummer Mitteilungen des Deutschen Germanistenverbandes: Korpora in der Linguistik – Perspektiven und Positionen zu Daten und Datenerhebung 1, 45–59.

Bubenhofer, Noah (2015): Kollokationen, n-Gramme, Mehrworteinheiten. In: Kersten Sven Roth/Martin Wengeler/Alexander Ziem (Hg.): Handbuch Sprache in Politik und Gesellschaft. Berlin / New York (Sprachwissen).

Bubenhofer, Noah (2013): Quantitativ informierte qualitative Diskursanalyse. Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: Kersten Sven Roth/Carmen Spiegel (Hg.): Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. Berlin (Diskursmuster - Discourse Patterns, 2), 109–134.

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin, New York (Sprache und Wissen, 4).

Bubenhofer, Noah/Nicole Müller/Joachim Scharloth (2013a): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. In: Zeitschrift für Semiotik, Methoden der Diskursanalyse 35 (3-4), 419–444.

Bubenhofer, Noah/Joachim Scharloth (2013a): Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In: Ingo Warnke/Ulrike Meinhof/Martin Reisigl (Hg.): Diskurslinguistik im Spannungsfeld von Deskription und Kritik. Berlin (Diskursmuster – Discourse Patterns, 1), 147–168.

Bubenhofer, Noah/Joachim Scharloth (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Daniel Elmiger/Alain Kamber (Hg.): La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretation. Neuchâtel (Travaux neuchâtelois de linguistique, 55), 241–259.

Bubenhofer, Noah/Joachim Scharloth (2013b): Korpuspragmatische Methoden für kulturanalytische Fragestellungen. In: Nora Benitt/Christopher Koch/Katharina Müller u.a.(Hg.): Kommunikation Korpus Kultur: Ansätze und Konzepte einer kulturwissenschaftlichen Linguistik. Trier (Giessen Contributions to the Study of Culture), 47–66.

Bubenhofer, Noah/Joachim Scharloth (2015): Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. In: Zeitschrift für Germanistische Linguistik 43 (1), 1–26.

Bubenhofer, Noah/Joachim Scharloth/David Eugster (2014): Rhizome digital: Datengeleitete Methoden für alte und neue Fragestellungen in der Diskursanalyse. In: Zeitschrift für Diskursforschung, Sonderheft Diskurs, Interpretation, Hermeneutik.

Bubenhofer, Noah/Juliane Schröter (2012): Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalyse. In: Péter Maitz (Hg.): Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. Berlin/Boston (Studia Linguistica Germanica, 110), 263–287.

Bubenhofer, Noah/Martin Volk/David Klaper u.a. (Hg.) (2013b): Text+Berg-Korpus (Release 147_v03). Abgerufen am 4.3.2014 von http://www.textberg.ch.

Busse, Dietrich/Wolfgang Teubert (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen

Semantik. In: Dietrich Busse/Fritz Hermanns/Wolfgang Teubert (Hg.): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen, 10–28.

Carstensen, Kai-Uwe/Christian Ebert/Cornelia Ebert u.a. (2010): Computerlinguistik und Sprachtechnologie. 3. Aufl. Heidelberg, Berlin Abgerufen am 14.01.2013 von http://www.springer.com/spektrum+akademischer+verlag/informatik/informatik+und+it+%C3%BCbergreifend/book/978-3-8274-2023-7.

Chen, Chun-houh/Wolfgang Härdle /Antony Unwin (Hg.) (2008): Handbook of data visualization. (Springer handbooks of computational statistics), Abgerufen am 4.3.2014 von http://sfx.ethz.ch/sfx_locater?sid=ALEPH:EBI01&genre=book&isbn=9783540330370&id=doi:10.1007/978-3-540-33037-0 Online via SFX.

Dunning, Ted (1994): Statistical identification of language.

DWB = Grimm, Jacob/Wilhelm Grimm (1854-1961): Deutsches Wörterbuch von Jacob und Wilhelm Grimm. 16 Bde. in 32 Teilbänden. Leipzig. Quellenverzeichnis Leipzig 1971, 2761–2771.

Evert, Stefan (2009): 58. Corpora and collocations. In: Anke Lüdeling/Merja Kytö (Hg.): Corpus Linguistics. Berlin/New York (Handbücher zur Sprach- und Kommunikationswissenschaft, 29), 1212–1248.

Feilke, Helmuth (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt am Main.

Feilke, Helmuth/Angelika Linke (Hg.) (2009): Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt. Berlin, New York.

Fleck, Ludwik: Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv, hg. von Lothar Schäfer und Thomas Schnelle, 10. Aufl., Suhrkamp Verlag 1980.

Ford, Paul (2015): What Is Code? If You Don’t Know, You Need to Read This. In: Businessweek Abgerufen am 13.03.2015 von http://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/.

Gabrilovich, Evgeniy/Shaul Markovitch (2006): Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI). Boston, 1301–1306.

Graham, Shawn/Scott Weingart/Ian Milligan (2012): Getting Started with Topic Modeling and MALLET. Abgerufen am 4.3.2014 von http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Ide, Nancy/Patrice Bonhomme/Laurent Romary (2000): XCES: An XML-based Encoding Standard for Linguistic Corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association.

Jung, Matthias (1996): Linguistische Diskursgeschichte. In: Karin Böke/Matthias Jung/Martin Wengeler (Hg.): Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Georg Stötzel zum 60. Geburtstag gewidmet. Opladen, 453–472.

Kath, Roxana/Gary S. Schaal/Sebastian Dumm (2015): New Visual Hermeneutics. In: Zeitschrift für germanistische Linguistik 43 (1), 27–51.

Keim, Daniel A./Jörn Kohlhammer/Geoffrey Ellis u.a. (2010): Mastering the information age - solving problems with visual analytics. Goslar. Abgerufen am 4.3.2014 von http://www.vismaster.eu/book/.

Kilgarriff, Adam (2001): Comparing Corpora. In: International Journal of Corpus Linguistics 6 (1), 1–37.

Krämer, Sybille (2009): Operative Bildlichkeit. Von der ‚Grammatologie' zu einer ‚Diagrammatologie'? In: Martina Heßler/Dieter Mersch (Hg.): Logik des Bildlichen. Zur Kritik der ikonischen Vernunft. Bielefeld (Metabasis, 2), 94–123.

Kruja, Eriola/Joe Marks/Ann Blair u.a. (2002): A short note on the history of graph drawing. In: Petra Mutzel/Michael Jünger/Sebastian Leipert (Hg.): Graph Drawing. (Lecture Notes in Computer Science, 2265), 272–286.

Kunze, Claudia/Lothar Lemnitzer (2002): GermaNet – representation, visualization, application. In: LREC. 1485–1491.

Kupietz, Marc/Cyril Belica/Holger Keibel u.a. (2010): The german reference corpus dereko: a primordial sample for linguistic research. In: Proceedings of the 7th conference on International Language Resources and Evaluation. Valletta, Malta, 1848–1854.

Lemke, Matthias/Alexander Stulpe (2015): Text und soziale Wirklichkeit. In: Zeitschrift für germanistische Linguistik 43 (1), 52–83.

Lemnitzer, Lothar/Heike Zinsmeister (2006): Korpuslinguistik. Eine Einführung. Tübingen.

Manovich, Lev (2015): Data Science and Computational Art History. In: International Journal for Digital Art History (1), 12–24.

Manovich, Lev (2014): Software is the message. In: Journal of Visual Culture 13 (1), 79–81, doi: 10.1177/1470412913509459.

Mautner, Gerlinde (2012): Corpora and Critical Discourse Analysis. In: Paul Baker (Hg.): Contemporary Corpus Linguistics. London/New York, 32–46.

Means, W. Scott (2004): XML in a Nutshell. Auflage: 3. Beijing ; Sebastopol, CA.

Miller, George A. (1995): WordNet: a Lexical Database for English. In: Communications of the ACM 38 (11), 39–41.

Mitchell, James Clyde (1969): Social networks in urban situations: analyses of personal relationships in central african towns. Manchester.

Newman, M. E. J. (2010): Networks: an introduction. Oxford/New York.

Perkuhn, Rainer/Holger Keibel/Marc Kupietz (2012): Korpuslinguistik. Stuttgart.

Pfeffer, Jürgen (2010): Visualisierung sozialer Netzwerke. In: Christian Stegbauer (Hg.): Netzwerkanalyse und Netzwerktheorie, 227–238.

Rieder, Bernhard/Theo Röhle (2012): Digital Methods: Five Challenges. In: David M. Berry (Hg.): Understanding Digital Humanities. Basingstoke, 67–84.

Rohrdantz, Christian/Annette Hautli/Thomas Mayer u.a. (2012): Towards tracking semantic change by visual analytics. Abgerufen am 04.03.2013 von http://kops.ub.uni-konstanz.de/handle/urn:nbn:de:bsz:352-186381.

Scharloth, Joachim/Noah Bubenhofer (2011): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Ekkehard Felder/Marcus Müller/Friedemann Vogel (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York, 195–230.

Scharloth, Joachim/David Eugster/Noah Bubenhofer (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Dietrich Busse/Wolfgang Teubert (Hg.): Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden, 345–380.

Schich, Maximilian/Chaoming Song/Yong-Yeol Ahn u.a. (2014): A network framework of cultural history. In: Science 345 (6196), 558–562, doi: 10.1126/science.1240064.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.

Schmid, Helmut/Florian Laws (2008): Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In Proceedings of the 22nd International Conference on Computational

Linguistics-Volume 1 (pp. 777-784). Association for Computational Linguistics.

Spitzmüller, Jürgen/Ingo H. Warnke (2011): Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin/New York.

Stegbauer, Christian (2010): Netzwerkanalyse und Netzwerktheorie : ein neues Paradigma in den Sozialwissenschaften. 2. Aufl. Wiesbaden (Netzwerkforschung).

Steinseifer, Martin (2013): Texte sehen – Diagrammatologische Impulse für die Textlinguistik. In: Zeitschrift für germanistische Linguistik 41 (1), 8–39.

Stjernfelt, Frederik (2007): Diagrammatology: an investigation on the borderlines of phenomenology, ontology, and semiotics. Dordrecht/London.

Storjohann, Petra/Melani Schröter (2011): Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise und die (Un-) Ordnung des Ausgeblendeten. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 7 (1), 32–53.

Stührenberg, Maik (2012): The TEI and current standards for structuring linguistic data. In: Journal of the Text Encoding Initiative (Issue 3), doi: 10.4000/jtei.523.

TEI Consortium (Hg.) (2014): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.6.0.

Teubert, Wolfgang (2006): Korpuslinguistik, Hermeneutik und die soziale Konstruktion von Wirklichkeit. In: Linguistik online 28 (3), 41–60.

Tidwell, Doug (2008): XSLT. Auflage: 2. Sebastopol, CA.

Tognini-Bonelli, Elena (2001): Corpus linguistics at work. Amsterdam (Studies in Corpus linguistics, 6).

van der Vlist, Eric (2011): XML Schema: The W3C's Object-Oriented Descriptions for XML. Auflage: 1.

Zesch, Torsten/Christof Müller/Iryna Gurevych (2008): Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Nicoletta Calzolari/Khalid Choukri/Bente Maegaard u.a. (Hg.): Proceedings of the Sixth International Language Resources and Evaluation (LREC 08). Marrakech, 1646–1652.

Zinsmeister, Heike (2015): Chancen und Grenzen von automatischer Annotation. In: Zeitschrift für germanistische Linguistik 43 (1), 84–111.